BASES: Large-scale Web Search User Simulation with Large Language Model based Agents

Anonymous ACL submission

Abstract

Due to the excellent capacities of large language models (LLMs), it becomes feasible to develop LLM-based agents for reliable user simulation. Considering the scarcity and limit (*e.g.,* privacy issues) of real user data, in this paper, we conduct large-scale user simulation for web search scenario to improve the analysis and modeling of user search behavior. Specially, we propose BASES, a novel user simulation framework with LLM-based agents, designed to facilitate comprehensive simulations of web search user behaviors. Our simulation framework can generate unique user profiles at scale, which subsequently leads to diverse search behaviors. To demonstrate the effectiveness of BASES, we conduct evaluation experiments based on two human benchmarks in both Chinese and English, demonstrating that BASES can effectively simulate large-scale human-like search behaviors. To further accommodate the research on web search, we develop WARRIORS, a new large-scale dataset encompassing web search user behaviors, including both Chinese and English versions, which can greatly bolster research in the field of information retrieval.

1 Introduction

Web search is a typical information-seeking scenario, where a user issues a query, retrieves the web pages from the search engine, and subsequently selects and browses through the pages of interest. The core to the success of search systems lies in the accurate understanding and modeling of user behavior. Thus, by analyzing large-scale user behavior under web search scenarios, we can enhance the understanding of user's information needs, thereby facilitating the development of more effective search systems (Bennett et al., 2012; Zhou et al., 2021; Chen et al., 2021). However, existing research works typically rely on real-user experiments (Carterette et al., 2014), which can be costly to scale up. Moreover, concerns regarding the quality and integrity of collected user data emerge as notable impediments, potentially undermining the precision of the analysis (Sugiyama et al., 2004). For example, web search users usually do not maintain a user profile, or the profile is incomplete or noisy. In addition, these experiments encompass ethical considerations, such as user privacy issues. As a result, user behavior simulation emerges as a promising direction to facilitate user behavior analysis and modeling (Jiang et al., 2012).

Recent years have witnessed the unprecedented success of large language models (LLMs) (Zhao et al., 2023). With superb model capabilities, LLMs are capable of comprehending complex instructions and executing actions as *autonomous agents* (Xi et al., 2023; Wang et al., 2023a). Prior studies have explored LLM-based agents in various fields, such as recommender systems (Wang et al., 2023b) and dialogue systems (Chan et al., 2023). In contrast with prior studies, we focus on a quintessential user interaction scenario of web search and utilize LLM-based agents to simulate web search user behaviors. Despite that this idea is intuitive, there exist several major challenges to simulate the behaviors of web search users. Firstly, each user is an independent individual, it is difficult to ensure that each user has a unique and reasonable profile, particularly at a large scale. Secondly, predicting web search user behaviors with precision and personalization is challenging.

To address these issues, we propose a novel framework **BASES** for large-scale weB seArch uSEr Simulation with LLM-based agents. First, we design user profile attributes tailored to the characteristics of web search users with reference to the related study (Brusilovsky et al., 2007). Each user profile encapsulates two basic categories of attribute types, static and dynamic, and a total of eight attribute types are allocated within the two categories. To construct unique profiles for large-scale simulated users, we propose a synergistic synthesis method, which predefines a comprehensive range of potential attribute values combining manual definition with the cooperation of GPT-4, ensuring both efficiency and diversity for user profile generation. Subsequently, LLM-based agents tailored with distinct user profiles are employed to conduct user simulations through the proposed query and click behavior prompting strategies in the search engine, thereby generating precise and personalized user behaviors.

To verify the effectiveness of the BASES framework, we first show the high anthropomorphism in search behaviors of the LLM-based agents by quantitative analysis, and conduct the manual evaluation on the simulated behavior data to demonstrate the personalization degree. Following these foundational assessments, we adopt BASES within both Chinese and English scenarios on two typical information retrieval (IR) tasks, including session search and click prediction. We find that models trained by user behavior data generated by BASES achieve improvements of up to 13% in NDCG@1 compared to other behavior datasets trained models, while utilizing less than one percent or one thousandth of their training data volume. Moreover, BASES also shows the considerable potential in the adaption of low-resource scenarios, bring up to 7% improvements in NDCG@1 with multiple adapting strategies. Finally, to accommodate the progressively envolving search patterns of users, we systematically collect the simulated data and construct a new dataset WARRIORS of Web seARch useR behavIORS with 100,000 users' search sessions for both Chinese and English versions. The key contributions are summarized as follows:

- We propose BASES, a novel web search user behavior simulation framework based on LLM-based agents, which is able to construct extensive and diverse user profiles and precisely perform user behaviors in web search scenarios.
- Extensive experiments on the proposed BASES framework across various IR tasks demonstrate that BASES significantly enhances the performance of IR models, and perform effectively in low-resource scenarios.
- We systematically collect and organize the large-scale simulated web search user behaviors, and release a new dataset WARRIORS, encompassing web search behaviors in both Chinese and English versions.

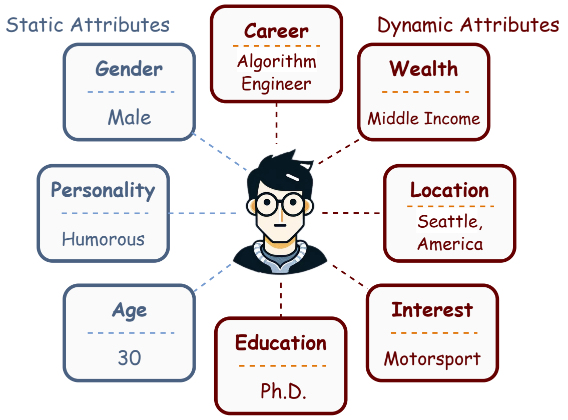

Figure 1: The profile structure for web search user simulation with a user sample.

2 Large-scale User Simulation for Web Search

In this section, we propose a novel framework for large-scale user Simulation in Web Search scenarios based on LLM-based agents, called **BASES**, with specially designed user profile construction strategies and tailored simulation process. We also discuss the superiority and feasibility of the simulation framework by evaluating the human-agent consistency.

2.1 User Profile Construction

In this part, we detail the process of user profile construction, including user profile structure, construction methods for user profiles, and the definition of attribute values.

2.1.1 User Profile Structure

To effectively simulate the search behaviors of web users, it is crucial to establish diverse and detailed user profiles tailored to web search scenarios. The profile should encompass various attributes that exert varying degrees of influence on search behaviors. For example, a younger user may exhibit a preference for searching and clicking on more trendy topics. Drawing on insights from prior research work (Brusilovsky et al., 2007; Liang, 2011), as shown in Figure 1, we design a profile structure of web search users, comprised of eight distinct attributes, including age, gender, education, career, personality, interest, location, and wealth, which are further categorized into static attributes and dynamic attributes. Static attributes are inherent to the user and tend to remain constant over time (*e.g.*, gender), while dynamic attributes reflect transient and external aspects and subject to change (*e.g.*, interest). It is essential for each simulated web search user to possess a unique profile, thereby enabling the simulation of a diverse range of search behaviors. Note that user profile designing is a classic and ongoing research topic in the community (Sugiyama et al., 2004; Heidari et al., 2020), while the core focus of our study is to demonstrate that LLM agents can be good web search user simulators. We consider the optimization of user profile settings as future work.

2.1.2 User Profile Construction Methods

To construct diverse and authentic user profiles, a basic method is involving human participants in filling out profile attributes. Despite its flexibility, it is inefficient and limited by the human annotator. Another intuitive method is utilizing LLMs such as GPT-4 (OpenAI, 2023) with instructions to automatically generate user profiles, but it is hard to ensure diversity.

To this end, we propose two profile construction methods: *Real Data based Generation* analyzes real user data with LLMs to generate the user profile. This method requires large scale of real user data and may be affected by the data quality. *Synergistic Synthesis* scientifically predefines attribute values through a collaborative effort involving human and GPT-4. These predefined attribute values are then randomly sampled based on various strategies, such as distribution patterns and logical coherence. Given the requirement to simulate web search users at a large scale, where both efficiency and diversity are important in profile construction, we adopt synergistic synthesis as the primary profile construction method.

With synergistic synthesis for user profile construction, we employ three principles for predefining attribute value candidates and sampling: (1) For attributes with easily-defined and uniformly-distributed candidate values (*e.g.,* gender), we manually define each attribute value and conduct random sampling. (2) For attributes with unclear value candidates and even distribution (*e.g.,* interest), we employ a coarse-to-fine sampling approach. First, we categorize attributes into various coarse-grained categories, subsequently integrating GPT-4 for the generation of fine-grained attribute values with manual adjustment within each category. (3) For attributes exhibiting non-uniform distribution of value candidates (*e.g.,* location), we consult pertinent literature to define potential attribute values, sampling them in accordance with their distribution. In addition, we meticulously identify and address the irrational aspects in profiles, including age-career contradictions, age-education contradictions, and career-education contradictions.

2.1.3 Attribute Value Definition

Here, we detail the specific definition and sampling methods employed in each attribute.

Age: In consideration of the authentic demographics of web search users, we implement random selections of user ages ranging from 6 to 90.

Gender: We implement random sampling to select user gender, including male and female.

Personality: We initially delineate the coarse-grained personality categories referring to the five-factor model (McCrae and John, 1992), including openness, conscientiousness, extraversion, agreeableness, and neuroticism. Subsequently, we define the fine-grained personality values with the collaboration of human and GPT-4 in each category. Finally, we obtain 128 personality values.

Education: Current statistics regarding the global population's educational levels are somewhat incomplete. Nevertheless, we refer to the reports from the UNESCO Institute for Statistics (UNESCO, 2023) and the World Bank (the World Bank, 2023) to approximately estimate the distribution of different educational levels of users.

Career: Referring to ISIC[1] from the United Nations, we initially delineate twenty coarse-grained career categories such as "Financial and insurance activities", student category and Non-working attribute are also included. Subsequently, we define the fine-grained career values with the collaboration of human and GPT-4 in each category. Finally, we obtain 2,743 attribute values in career.

Interest: Collaborating with GPT-4, we initially delineates eleven coarse-grained interest categories such as "Sports and outdoor activities". Subsequently, we define the fine-grained interest values with the collaboration of human and GPT-4 within each coarse-grained category. Finally, we obtain 1,618 attribute values in interest.

Location: We consult global demographic data[2] in 2023 to identify the top 1,000 cities ranked by population in English setting, serving as candidate locations. For the Chinese setting, we use cities in China with populations as candidates (Ning, 2021).

---

[1] `https://unstats.un.org/unsd/classifications/Econ/isic`

[2] `https://worldpopulationreview.com/world-cities`

Then, we randomly sample cities weighted according to the population distribution of the cities.

Wealth: We categorize users into three distinct wealth status groups: low income, middle income, and high income. Referring to the existing studies (Kochhar, 2021), we randomly sample each wealth category weighted by corresponding approximate population proportion.

2.2 Web Search User Simulation

Based on the aforementioned profile construction method, we employ LLM-based agents in search engine environments to conduct large-scale user behavior simulation for web search.

2.2.1 Simulation Essentials

In our BASES framework, the agents for user simulation are required to be aware of their current state in the search scenario, plan to proactively formulate search queries, assimilate information returned by search engines, and selectively click web pages that match their interests. Additionally, the agents need to reflect whether their information needs have been satisfied based on the retrieved information, to decide whether new queries are warranted for further searches.

Although existing studies have successfully utilized LLM-based agents for simulating and fulfilling particular tasks under different environments (Park et al., 2023; Wang et al., 2023b) , the capacity of LLM-based agents to accurately mimic real users in web search scenarios remains unclear. We evaluated the consistency of behavioral patterns between agents and real users within the search engine, focusing on fundamental search actions, to prove the feasibility of web search user simulation. We find that LLM-based agents show about 90% accuracies for query behaviors consistency and click behaviors consistency. The details of the assessment can be found in Appendix A.

2.2.2 Simulation Process

Behavior Patterns. We employ the web search user profiles constructed using synergistic synthesis described in Section 2.1.1, each user profile corresponding to an LLM-based agent for user simulation. We utilize the ChatGPT with `GPT-3.5-Turbo-1106` version as the LLM to implement the web user agents. These agents, acting on behalf of their respective users, engage in multi-round interactions in search sessions with search engines. We define three types of actions: search, click, and finish. In each round, an agent first generates a query based on its user profile and historical behaviors, which is then submitted to the search engine and the search engine returns relevant web pages from the internet. Subsequently, the agent selects web pages of interest to click on, based on its profile, historical behavior, the current query, and the returned web pages. At the end of each round, the agent automatically decide whether to continue with another round or to finish the current search session.

Behavior Prompting Strategies. An intuitive idea for simulating user behavior with LLM-based agents, similar to ReAct (Yao et al., 2023), involves defining all possible actions within one instruction template, planning the overall steps before task execution, or thinking progressively during the task execution. We make such an attempt in the web search scenario and find that predefining all possible actions within one instruction is not beneficial for the agents to effectively mimic human user behavior. This issue is primarily manifested in the arbitrary click behaviors of the agents.

Consequently, we propose two instruction strategies to prompt agents to execute precise actions, including query behavior prompting and click behavior prompting. *Query behavior prompting* is tailored to enable the agents to generate queries for web search based on their user profiles and historical behaviors. Considering the search pattern of real users, we limit the queries to be concise, primarily focusing on keyword keyphrases. The decision to finish or continue the session is also integrated, since the search action is the first action of each round. If the agent opts to continue the session, it performs a search action to start a new round, otherwise, it executes a finish action. *Click behavior prompting* is designed to guide the agent in selecting web pages to view from search engine results, considering its user profile, current query, and historical behaviors. We fully consider real web search scenarios, providing agents with the top ten web pages returned by search engines. When clicking, the agent primarily rely on the web page titles, and not only select the webpage IDs but also to provide explanations for their choices, so as to prevent arbitrary clicks.

For various prompting strategies, we design specific templates to ensure the outputs of agents are stable and controllable. The full instruction templates are reported in Appendix C. Finally, each user exhibits sequential behaviors (*i.e.,* a search session) within the web search scenario. We show an example of our simulated search user behaviors in Appendix D.

2.3 Discussion

With BASES framework, it is feasible to simulate highly personalized and diverse search behaviors via LLM-based agents. We propose a profile construction method that is capable of effortlessly generating large scale distinct user profiles. Through meticulously constructed user profiles, BASES can efficiently and effectively produce the behaviors for these users through two behavior prompting strategies, and the behaviors generated by BASES can reflect user biases through the discrepancy between different user profiles. It greatly benefits the application scenarios where real user data is limited or difficult to collect. Moreover, given the controllability of the simulated web search users, our BASES framework can flexibly incorporate various constraints to meet the further demands of specific tasks, such as conducting long search sessions, formulating queries in unique formats, and supporting distinct rules for clicks, among others.

Personalization is critical in web search behavior simulation, its absence can lead to homogenization in user behaviors, hindering subsequent tasks. We manually assess the personalization in the simulated user behaviors of BASES, and find that even users with extremely similar profiles exhibited significantly inconsistent behavior patterns. The details can be found in Appendix B.

3 Effectiveness Verification

In this section, we investigate the effectiveness of our proposed BASES framework on classical information retrieval (IR) tasks using evaluation benchmarks constructed with real user data.

3.1 Web Search User Simulation for IR tasks

3.1.1 Setups

Task Formulation. In general, the BASES framework can be applied to various IR tasks, among which we consider two classical IR tasks: session search and click prediction. *Session search* leverages the rich contextual information (such as queries $\{q_i\}_{i=1}^{n-1}$ and click history $\{d_i\}_{i=1}^{n-1}$) from earlier rounds of a user's session to aid document ranking for the subsequent search query q_n. It aims to enhance the learning of user preferences, thereby predicting the document d_n most aligned with the user's intent from candidate documents $\{d_n^j\}_{j=1}^{k}$. The task of *click prediction* aims to predict the documents most relevant to the current search query q from the search results $\{d_j\}_{j=1}^{k}$.

Evaluation Model. We employ a BERT-based ranking model for evaluation on one NVIDIA GeForce RTX 3090, training up to 3 epochs with a batchsize of 128 and a learning rate of 5e-6. For the session search task, the model's input is the concatenation of historical user behavior sequence $\{q_i, d_i, ..., q_{n-1}, d_{n-1}\}$, the current query q_n, and the candidate document d_n^j. The output of the model is a relevance score between the document d_n^j and the current query with behavior history. For the click prediction task, the model's input is the concatenation of query q and candidate document d_j, with the output being the relevance score between the query q and the candidate document d_j.

Evaluation Metrics. The evaluation metrics utilized in our study include Mean reciprocal rank (MRR) and Normalized Discounted cumulative gain (NDCG). MRR quantifies search accuracy by averaging the reciprocal ranks of the first relevant result across queries. NDCG assesses ranking quality by comparing the weighted relevance of all results to an ideal order, we use NDCG@1 and NDCG@3 for evaluation.

3.1.2 Evaluation on Real User Behavior Benchmarks

To show the efficacy of BASES, we employ it to generate behavior data of simulated web search users in both Chinese and English settings and train user behavior models. We collect *search behaviors from real users* and construct benchmarks on two classic IR tasks for evaluation.

Evaluation Benchmark. In the Chinese setting, we randomly sampled 100 anonymous sessions from recent search logs obtained via Baidu Search, with each session encompassing the user's historical queries, search results, and click records. Since recent English user search behavior data is not available, we develop an English benchmark through the collaboration of GPT-4 and humans. Initially, we instruct GPT-4 to produce query sequences similar to those generated by real users and obtain a collection of 100 diverse query sequences. Human annotators are then involved in clicking one of the search results considering history behavior

Tasks	Methods	#Session (#Click)	*Chinese Benchmark*			*English Benchmark*		
			MRR	NDCG@1	NDCG@3	MRR	NDCG@1	NDCG@3
Session Search	BM25	-	45.16	27.20	41.39	33.06	14.43	25.39
	BERT (TREC-Session)	1,257	-	-	-	32.85	13.11	25.89
	BERT (AOL)	219,748	-	-	-	35.59	16.07	29.54
	BERT (Tiangong-ST)	143,155	43.28	22.59	38.91	-	-	-
	BERT (BASES)	1,000	51.78	35.56	47.98	39.59	19.02	34.22
	BERT (BASES)	10,000	**53.52**	**35.98**	**50.72**	**40.86**	**20.33**	**35.50**
Click Prediction	BM25	-	42.71	21.34	39.96	31.35	11.80	23.88
	BERT (TREC-Session)	1,654	-	-	-	32.32	12.13	25.89
	UBM (AOL)	614,651	-	-	-	24.28	10.03	19.69
	BERT (AOL)	614,651	-	-	-	33.71	15.41	26.94
	BERT (Tiangong-ST)	318,823	41.29	21.34	36.28	-	-	-
	BERT (MS MARCO)	502,193	-	-	-	34.99	16.07	27.49
	BERT (ORCAS)	88,689,558	-	-	-	36.73	16.12	30.82
	BERT (BASES)	2,880/2,920	46.24	27.20	42.39	36.74	16.07	30.37
	BERT (BASES)	28,511/28,802	**49.11**	**30.13**	**45.92**	**38.97**	**18.69**	**32.13**

Table 1: Results of methods trained on different datasets in real user benchmarks across two IR tasks.

and current query to construct the search sessions.

Baselines. Since the our primary objective is to demonstrate the effectiveness of our simulated behavior data, the comparisons focus on assessing the performance of models trained on our simulation data versus various existing datasets, all under the same model backbone, on real-world test sets. For comparison, we select three widely adopted session search datasets for model training, including TREC-Session (Carterette et al., 2014) and AOL (Verbeek and Slob, 2006) for the English benchmark, and Tiangong-ST (Chen et al., 2019) for the Chinese benchmark. We also apply above datasets for click prediction task, and adopt ORCAS (Craswell et al., 2020) and MS MARCO (Nguyen et al., 2016) as click prediction datasets. We train BERT-based ranking models (as described in Section 3.1.1) on the three datasets respectively. In addition, we also adopt a classic click model UBM (Dupret and Piwowarski, 2008) for comprehensive comparison.

Results and Analysis. Table 1 reports the results of various methods on benchmarks constructed with real user behaviors. We observe that the models trained using user behavior data constructed by BASES significantly outperform the baselines in both session search and click prediction tasks, even with a small amount of training data. We also find that BERT-based click model trained with ORCAS achieves the best performance among all baselines with more than 88 million training data, while we achieve comparable performance using only over 2,000 simulation data, with additional improvements observed upon further increasing the data volume. The results prove that BASES can accurately simulate web search users performing search processes in search engines like real users, and the simulation data significantly enhances the performance of information retrieval models. Moreover, we find that the classic lexical model BM25 shows considerable competitiveness, surpassing some semantic models trained on web search behavior datasets, and the traditional click model shows lower performance comparing to the BERT-based model with the same training set. Furthermore, as the performance of search engines has improved in recent years, which are able to handle more complex queries, the query formats made by web search users are gradually influenced, leading to discrepancies from the query formats in previous datasets. Therefore, our simulation methodology is of significant importance for generating behaviors that align with contemporary human search habits.

Training Data	#Session	MRR	NDCG@1	NDCG@3
Origin	800	60.89	40.65	53.03
Ori. + Desc.	1600	**65.28**	**47.97**	57.27
Ori. + Behav.	1600	65.20	46.34	**57.44**
Ori. + Synth.	1800	62.80	45.53	52.86
Ori. + Synth.	2300	62.37	43.09	54.27
Ori. + Synth.	2800	63.85	46.34	55.00

Table 2: Results of BASES-augmented methods on TREC-Session. *Origin* (*Ori.*) denotes the training data of TREC-Session. *Desc.*, *Behav.*, and *Synth.* denote the training data constructed by BASES with profiles generated by descriptions and user behaviors in TREC-Session, and synergistic synthesis, respectively.

3.2 Adaptation on Low-resource Scenarios

Acquiring large-scale real user behavior data presents significant challenges. In this section, we apply BASES framework to improve the performance of low-resource web search scenarios.

Experimental Setup. We consider TREC-Session for experiments. Since it is annotated by human participants, the overall volume of data is not substantial, which is a typical low-resource web search scenario. In this scenario, we apply our BASES framework to generate augmented data on the session search task. We construct user profiles for simulation using methods in Section 2.1.2. The user profile are generated with ChatGPT using real data based generation by employing descriptions (**Desc.**) or user behaviors (**Behav.**) in the dataset, and using the method of synergistic synthesis (**Synth.**) without referring to the dataset. We also conduct experiments with various scales of profiles generated by Synth.. Subsequently, we utilize the three kinds of profiles to simulate augmented user behaviors for training. We use the same BERT-based ranking models for each settings.

Results and Analysis. Table 2 presents the results on TREC-Session employing various augmentation methods with BASES. It can be observed that all the methods with augmented data can yield improvements over the baseline established by training on the original TREC-Session training set. This demonstrates the effectiveness of the BASES framework to address the data scarcity issue in low-resource web search scenarios, thereby augmenting the original behavioral data to secure improved performance. Furthermore, augmentations with Synth. achieve improvements, with the degree of improvement exhibiting an ascending trend in correlation with the amount of augmented data. This phenomenon emphasizes the ability of BASES to bring substantial benefits in low-resource scenarios without necessitating dataset-specific information.

4 The WARRIORS Dataset

Previous experiments verified the effectiveness of the BASES framework. To this end, we construct **WARRIORS**, a large-scale simulation dataset of Web seARch useR behavIORS based on BASES.

4.1 Existing Issues

For web search tasks, it is crucial to construct high-quality user behavior datasets. In the literature, several datasets have been proposed and contribute significantly to the advancement of the field. However, these datasets come with certain limitations. For instance, the AOL dataset (Pass et al., 2006) is 18 years old now, where many URLs are no longer accessible. TREC-Session (Carterette et al., 2014) is constrained by crowdsourced experiments, lacking large-scale records of user behavior. Tiangong-ST (Chen et al., 2019) is a large-scale web search behavior dataset in Chinese, defining session boundaries based on natural pauses of 30 minutes in user activity, which may introduce errors in session demarcation.

Dataset Statics	Chinese	English
#user	100,000	100,000
#avg. query per user	2.85	2.88
#avg. click per query	1	1
#avg. click position	2.6	4.1
Search engine	Google Search	
Time	Feb., 2024	

Table 3: Statistics of the WORRIORS dataset.

Furthermore, web search user behaviors are highly affected by the underlying algorithms of search engines. As the performance of search engines has gradually improved, the characteristics reflected in user behaviors within search engines have also evolved. For example, in the era of search engines reliant on keyword matching, users typically conducted searches using a single word or very short phrases. However, in recent years, with the rapid development of semantic search technologies (Zhao et al., 2024), user queries have shifted towards longer phrases and even sentences, as these approaches enable more efficient retrieval of the desired information. This shift implies that historical datasets of web search user behaviors may no longer reflect current user behaviors, presenting challenges for applying these datasets to enhance contemporary web search experiences.

Given the capability of BASES, it can be naturally used to construct a new dataset of web search behaviors using the data generated by the anthropomorphic users simulated by LLM-based agents.

4.2 Construction of WARRIORS

The construction of the WARRIORS dataset follows the BASES framework with LLM-based agents introduced in Section 2. The WARRIORS dataset is comprised of two versions: WARRIORS-Chinese and WARRIORS-English. For both versions, we use Google Search as the search engine.

Tasks	**Methods**	***WARRIORS-Chinese***			***WARRIORS-English***		
		MRR	**NDCG@1**	**NDCG@3**	**MRR**	**NDCG@1**	**NDCG@3**
Session Search	BM25	31.15	12.15	23.56	31.98	11.94	25.11
	BERT (TREC-Session)	-	-	-	33.33	13.78	26.60
	BERT (AOL)	-	-	-	34.85	14.70	28.24
	BERT (Tiangong-ST)	34.40	13.91	27.88	-	-	-
	BERT (WARRIORS)	**59.40**	**39.73**	58.97	46.96	27.29	43.32
	COCA (WARRIORS)	59.11	39.20	**58.99**	**47.07**	**27.78**	**43.39**
Click Prediction	BM25	30.44	11.62	23.09	32.10	12.12	25.20
	BERT (TREC-Session)	-	-	-	34.03	13.93	27.62
	BERT (AOL)	-	-	-	31.94	11.84	24.79
	BERT (Tiangong-ST)	34.85	15.57	28.31	-	-	-
	BERT (ORCAS)	-	-	-	36.74	16.12	30.82
	BERT (WARRIORS)	**57.49**	**38.01**	**56.73**	**46.68**	**27.00**	**43.20**

Table 4: Evaluation of several baselines on WARRIORS-Chinese and WARRIORS-English.

The user profiles of the two versions are independently generated, without overlapping users.

In search engines, the user behavior is characterized by the search session. Each user session encompasses multiple interaction rounds, each comprising a query, search results from the query, and clicked web pages from search results. The search results include the top ten web pages returned by the search engine, each web page with its URL, title, and snippet. Appropriate constraints are increased on the instructions for LLM agents, such as limiting the number of interaction rounds to avoid extremely long sessions. The entire dataset is further divided into training set, validation set, and test set. Given the substantial volume of the dataset, the specific allocation was 98% for the training set, with the validation and test sets each receiving 1%. Table 3 shows the details of the dataset.

4.3 Dataset Application

Experiment Setup. WARRIORS can be easily applied to a variety of IR tasks, such as session search, click prediction, query suggestion, and so on. Here, we implement and evaluate several baselines on session search and click prediction tasks with both WARRIORS-Chinese and WARRIORS-English. For each task, we apply two representative models, a lexical matching model BM25 (Lin et al., 2021) and a BERT-based (Devlin et al., 2019) semantic ranking model. Furthermore, we report the results of a BERT-based ranking model trained using existing web search user behavior datasets (Chen et al., 2019; Carterette et al., 2014) for reference.

Results and Analysis. Table 4 reports the results of the baselines on WARRIORS-Chinese and WARRIORS-English. we find that the BERT-based models trained on our dataset achieve the best performance on both session search and click prediction tasks, surpassing BM25 and models trained on existing datasets in both Chinese and English settings. The results demonstrate the capacity of our dataset to assess various models and IR tasks. It can be seen that WARRIORS provides an ideal foundation for enhanced IR models within web search scenarios, containing near-real user interaction behaviors and the latest web search results.

5 Conclusion

In this study, we introduced BASES, a new framework designed for the simulation of large-scale web search user behaviors with LLM-based agents. We designed specific strategies to construct large-scale user profiles, which can subsequently lead to human-like search behavior data in our framework. Through extensive experiments, we not only demonstrated the effectiveness of BASES in simulating authentic and diverse user behaviors, but also highlighted its potential in improving information retrieval tasks, especially in low-resource scenarios. Furthermore, to facilitate the related research, we develop the WARRIORS dataset, available in both Chinese and English versions, which comprises a large collection of simulated web search user behavior data by BASES. We believe that this work can provide new perspectives to investigate the search behaviors of web users, thereby contributing to the advancement of search technologies and user experience optimization. For future work, we will enrich the behavior types of simulated users, and also consider applying the BASES framework to other fields like sociology analysis.

Ethical Considerations

Given the limitations and privacy concerns associated with real user data, we introduce BASES, a framework for simulating large-scale web search user behaviors using LLM-based agents. BASES demonstrates promising potential in enhancing information retrieval tasks, especially in low-resource scenarios. This approach not only respects user privacy but also opens up new avenues for research in this domain, contributing positively to the advancement of information retrieval technologies. As we move forward, we are committed to ensuring that our simulation practices uphold the ethical standards and contribute constructively to the field.

Limitations

Our user simulation methodology exhibits highly realistic simulated behaviors and good application performance. However, due to the complexity of human, the user profiles we constructed do not comprehensively encapsulate all human characteristics. Although minor, this fact also harbors the potential to influence search behaviors of users. In the future, we will continue to explore schemes for simulating web search users, aiming to align simulated behaviors as closely as possible with those of real users.

References

Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short- and long-term behavior on search personalization. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 185–194. ACM.

Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors. 2007. *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*. Springer.

Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. 2014. Overview of the TREC 2014 session track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, volume 500-308 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201.

Jia Chen, Yiqun Liu, Jiaxin Mao, Fan Zhang, Tetsuya Sakai, Weizhi Ma, Min Zhang, and Shaoping Ma. 2021. Incorporating query reformulating behavior into web search evaluation. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 171–180. ACM.

Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2485–2488. ACM.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2983–2989. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 331–338.

Maryam Heidari, James H Jones, and Ozlem Uzuner. 2020. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 480–487. IEEE.

Jiepu Jiang, Daqing He, Shuguang Han, Zhen Yue, and Chaoqun Ni. 2012. Contextual evaluation of query reformulations in a search session by user simulation. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 2635–2638. ACM.

Rakesh Kochhar. 2021. Are you in the global middle class? find out with our income calculator. Https://www.pewresearch.org/short-reads/2021/07/21/are-you-in-the-global-middle-class-find-out-with-our-income-calculator/.

Chunyan Liang. 2011. User profile for personalized web search. In *Eighth International Conference on Fuzzy*

Systems and Knowledge Discovery, FSKD 2011, 26-28 July 2011, Shanghai, China, pages 1847–1850. IEEE.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Frassetto Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2356–2362. ACM.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*.

Jizhe Ning. 2021. Main data of the seventh national population census. *National Bureau of Statistics of China*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006*, volume 152 of *ACM International Conference Proceeding Series*, page 1. ACM.

Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. 2004. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 675–684. ACM.

the World Bank. 2023. Education statistics (edstats). Https://gem-report-2023.unesco.org/.

UNESCO. 2023. 2023 global education monitoring report. Https://gem-report-2023.unesco.org/.

Peter-Paul Verbeek and Adriaan Slob. 2006. *User behavior and technology development*. Springer.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023a. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432.

Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. *CoRR*, abs/2306.02552.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Yujia Zhou, Zhicheng Dou, Bingzheng Wei, Ruobing Xie, and Ji-Rong Wen. 2021. Group based personalized search by integrating search behaviour and friend network. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 92–101. ACM.

A Simulation Consistency Evaluation

To demonstrate the feasibility of web search user simulation, we assess the consistency between the behavioral patterns exhibited by LLM agents and those of real-world users within the search engine.

A.1 Settings

Profile Generation. We conduct the experiment on TREC Session (Carterette et al., 2014), a human-annotated session search dataset encompassing user sessions of query-search-click behaviors, aimed at fulfilling informational needs of given task descriptions. Due to the limited volume of the dataset, we use ChatGPT as LLM-based agents to conduct real data based profile generation method in Section 2.1.2 with existing task descriptions.

Evaluation Tasks. Given the impracticality of evaluating the consistency of entire behavioral sequences, we deconstruct them and assess the consistency from two distinct behaviors: query and click. The query behavior evaluation including query generation and query rewriting. Query generation entails assessing the consistency between the initial queries generated by LLM agents, and those produced by humans. Query rewriting involves evaluating LLM agents' capabilities in formulating the subsequent query based on prior behaviors, ensuring it aligns with the human-rewritten query. The click behavior evaluation focuses on the accuracy of click behaviors. It evaluates the consistency of LLM agents' selections in search results based on historical behaviors with those of humans.

Evaluation Metrics. We use two metrics to evaluate query behaviors: term overlap rate and GPT-4-based evaluation. The term overlap rate quantifies the extent of shared subject keywords between paired queries. This is determined by identifying common words after removing stop words, since shared terminology indicates thematic similarity and a consistent search intent. Despite its simplicity and efficacy, term overlap rate has limitations in overlooking synonymous terms and varying morphological forms. GPT-4-based evaluation is an supplement, which assesses the consistency of the query generated by the LLM agent with the human-query's search intent and its adherence to the user profile. To evaluate click behaviors, we utilize top-1 accuracy to measure the proportion of instances where the LLM agent's clicked web page correspond to the same click of the human.

A.2 Results and Analysis

For query generation, the term overlap rate between queries generated by LLM agents and those produced by humans is found to be 90.8%, accompanied by the GPT-4 evaluation of 99%. For query rewriting, rewritten queries from LLM agents show a term overlap of 82.5% with human-rewritten queries, accompanied by the GPT-4 evaluation of 98.8%. The findings suggest that the LLM agents are capable of generating queries that align closely with human search objectives, both in scenarios involving and lacking historical interactions. Regarding the click prediction, the top-1 accuracy of the LLM agents' clicks mirroring human clicks is 51%. Although this value seems not so high, but in the industry, achieving such a top 1 accuracy is considered high, as many search services struggle to reach above 20% top 1 accuracy with their click models. Moreover, given the stringent nature of this metric, we manually review cases where the LLM agents do not precisely match the web pages clicked by humans. We find that in over 90% of these cases, the web pages clicked by the LLM agents have minor discrepancy with those of humans and consistent with user profiles. This indicates that click behavior is influenced by the homogenization of search results and individual bias in human participants, and fundamentally, the LLM agent's click behavior aligns well with human patterns.

Overall, across both dimensions of query and click, LLM agents demonstrate a high degree of consistency with human behaviors, thereby ensuring the viability of user simulation.

B Personalization Evaluation

Whether user behaviors are personalized is a critical evaluation factor in user simulation, especially when simulating large-scale web search users. If the simulation lacks personalization, it can lead to homogenization in the behavior of simulated users, meaning similar users are likely to perform the same search actions, which is detrimental to subsequent tasks.

Therefore, we focus on assessing the personalization in the simulated user behaviors of BASES framework. We construct a series of simulated user pairs, which demonstrate high similarity across multiple attributes in their profiles. For each pair of similar users, we manually analyze the behavior data generated by our BASES framework, the

User Profile	**Query Sequence**
"Career": "Performance Poet", "Age": 64, "Gender": "Female", "Education": "No formal education", "Location": "Delhi, India", "Wealth": "Middle income", "Personality": "*Amiable*", "Interest": "Ethnic cuisine cooking"	Indian vegetarian recipes, traditional Indian curry recipes, Indian street food recipes.
"Career": "Performance Poet", "Age": 64, "Gender": "Female", "Education": "No formal education", "Location": "Delhi, India", "Wealth": "Middle income", "Personality": "*Sociable*", "Interest": "Ethnic cuisine cooking"	cooking clubs or groups in Delhi, Traditional Indian cooking classes in Delhi, Indian cooking festivals.

Table 5: An example of how slight modifications in the user profile can lead to changes in user search behavior.

Cohen's Kappa of human participants is 0.9. We find that even users with extremely similar profiles (differing in only 1 or 2 attributes) exhibited significantly inconsistent behavior patterns, Table 5 gives an example, it can be observed that the two user profiles differ in only personality attribute, yet result in distinctly different simulated user behaviors. The evaluation result proves that the user behaviors simulated with BASES framework possess a significant degree of personalization. Furthermore, it also demonstrate that BASES framework provides deep insights into the complexity and diversity of user behavior patterns.

C Prompt Design

To prompt agents for executing precise actions, we devise two instruction strategies, including query behavior prompting and click behavior prompting, the specific prompt templates are shown in Table 6 for English setting and Table 7 for Chinese setting.

D Simulated Search User Behavior Example

We show an example of the simulated search user behavior.

Query Behavior Prompting

You are a search engine user with your own profile. Your task is to interact with search engines {max_exceeds_times} times. You have two types of operations to perform:

- Search[query]: When the current round has not reached {max_exceeds_times} times, please raise the next question based on your profile and web browsing history. Your query should be an entity phrase that has a similar topic to an attribute in your user profile. The query must be concise and clear. For example, Search[bollywood growth], Search[hjunk food trax], Search[pseudocyesis information], Search[location of port arthur].
- Finish[finish], When the current round exceeds {max_exceeds_times} times, you need to end your interaction with search engines. For example, Finish [Finish]

** Your Profile ** {profile}

** Web browsing (click) history ** {scratchpad}

** Your action **

Click Behavior Prompting

You are a search engine user with your own profile. Your task is to click on the most relevant page.
In this interaction, you raised the question of **query**. You have received several webpage titles returned by the search engine.
Now, based on your profile, web browsing history, and the relevance between the query and the titles, please choose the most relevant webpage to click on. Please note that you can only output one number from 1 to 10 to represent the title you are about to click on, and cannot output any other content

** Your Profile **
{profile}

** Web browsing (click) history **
{scratchpad}

** Query **
{query}

** Titles **
{titles}

** Your click **

Table 6: Various prompting strategies with query behavior prompting and click behavior prompting in English.

Query Behavior Prompting

您是一位带有自己档案的搜索引擎用户。您的任务是与搜索引擎互动{max_exceeds_times}次。您有两种操作可以执行：
- Search[查询]：在当前轮次未达到{max_exceeds_times}次时，请同时参考网页浏览历史和档案提出下一个查询。你需要考虑的是：1、在参考网页浏览历史时，查询的主题不能过多偏离历史中的交互内容，保持主题的连贯性；2、在参考档案时，查询的主题要和档案中的某属性相关，但要有发散思维，不要限制在这些具体属性上，尤其注意查询不要过多的涉及地名。
最后，您的查询应该是一个实体短语，保持简洁明了。这里有一些该操作的例子，如Search[优酷客户端下载], Search[无敌铁桥三], Search[蜡笔小新], Search[百度云网盘资源]。
- Finish[finish]：在当前轮次超过{max_exceeds_times}次时，您需要结束与搜索引擎的互动。例如，Finish[finish]。

** 您的档案**
{profile}

** 网页浏览（点击）历史**
{scratchpad}

** 您的操作**

Click Behavior Prompting

您是一位带有自己档案的搜索引擎用户。您的任务是点击最相关的页面。
在这次互动中，您提出了**query**的查询。您已经收到了搜索引擎返回的几个网页标题。
现在，请根据您的档案、网页浏览历史以及查询和标题之间的相关性，选择最相关的网页进行点击（语义相关或者是词语相似度高）。请注意，请首先输出你选择该title的理由，然后输出一个**数字**（1到10）来代表您即将点击的标题，例如我会选择Title x: xxxx, 因为xxx。

** 您的档案**
{profile}

** 网页浏览（点击）历史**
{scratchpad}

** 查询**
{query}

** 标题**
{titles}

** 您的点击**

Table 7: Various prompting strategies with query behavior prompting and click behavior prompting in Chinese.

User behavior data example:

```
{
"session": 3,
   "profile": {"Career": "Pricing Actuary", "Age": 22, "Gender": "Female", "Education": "Master",
        "Location": "Barcelona, Spain", "Wealth": "Low income", "Personality": "Cooperative",
        "Interest": "English Garden Design"}},

    "interaction": [
        {
        "step": 1,
        "action": {"type": "Search", "key": "pricing actuary job opportunities in Barcelona"},
        "observation": [{"title": "Actuarial Jobs in Barcelona | ProActuary", "snippet": "POPULAR
         JOBS. Actuary Jobs (1896) Management Jobs (715) Actuarial Analyst Jobs (514)...", "url":
            "https://proactuary.com/barcelona"}, . . .  ],
        "click": ["2"]
        },
        {
        "step": 2,
        "action": {"type": "Search", "key": "actuarial pricing training courses in Barcelona"},
        "observation": [{"title": "Training Courses in Barcelona, Spain - Oxford Management Centre",
            "snippet": "Our carefully curated training courses in Barcelona cater to...", "url":
            "https://oxford-management.com/training-venue/barcelona"}, . . . ],
        "click": ["1"]
        },
        . . .
    ]
}
```

Table 8: An example of search user behavior data generated by BASES in json format.