Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA 2023), a Satellite Workshop of the NTCIR-17 Conference, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001350

Fairness-based Evaluation of Conversational Search: A Pilot Study

Tetsuya Sakai
Waseda University, Japan
tetsuyasakai@acm.org

ABSTRACT

NTCIR-17 introduced the FairWeb-1 task, which evaluated web page rankings in terms of both relevance and group fairness. The present study shows how their evaluation framework can be extended for the evaluation of multi-turn, textual conversational search systems. By using the full test topic set of FairWeb-1 to harvest actual user-system conversations from the New Bing and Google Bard, we demonstrate how a series of system turns can be evaluated using our evaluation framework, which we call GFRC (Group Fairness and Relevance of Conversations). In addition, based on observations from our pilot experiment, we briefly discuss a few open questions in human-in-the-loop evaluation of conversational search in general.

1 INTRODUCTION

For the past few decades, offline web search evaluation usually meant "evaluating a ranked list of URLs" with measures such as nDCG (normalised Discounted Cumulative Gain) [12] and ERR (Expected Reciprocal Rank) [6]. However, the advent of conversational search engines based on large language models (LLMs) is rapidly changing the research landscape in search: in various search scenarios, the user may prefer to receive direct answers from a conversational search engine rather than a list of URLs from a traditional one. Evaluating a series of textual responses from conversational search is of utmost importance, as while their fluency may mislead the user into thinking that their responses are trustworthy and fair, they in fact *hallucinate* often, and may be biased, or even be harmful: see, for example, Askell et al. [1], Liang et al. [14], Liu et al. [16], Sakai [21].

Within the traditional ranked list evaluation paradigm, the NTCIR-17 FairWeb-1 task [26][1] evaluated participating runs based not only on relevance (for the benefit of the searcher) but also on *group fairness* [7] (for the benefit of the items being ranked or their stakeholders); to this end, they employed the GFR (Group Fairness and Relevance) framework [25]. The present study shows how this framework can be extended for the evaluation of multi-turn, textual conversational search systems. By using the full test topic set of FairWeb-1 to harvest actual user-system conversations from the New Bing[2] and Google Bard,[3] we demonstrate how a series of system turns can be evaluated using our evaluation framework, which we call GFRC (Group Fairness and Relevance of Conversations). In addition, based on observations from our pilot experiment, we briefly discuss a few open questions in human-in-the-loop evaluation of conversational search in general.

[1]http://sakailab.com/fairweb1/
[2]https://www.bing.com/
[3]https://bard.google.com/

2 RELATED WORK

2.1 GFR (Group Fairness and Relevance)

Given M different attribute sets (i.e., sets of groups) to consider for group fairness, the GFR score for evaluating a ranked list (L) of documents is defined as follows [25].

$$GFR(L) = \sum_{k=1}^{|L|} Decay(L,k) \left(w_0 Utility(L,k) + \sum_{m=1}^{M} w_m DistrSim^m(L,k) \right) . \tag{1}$$

Here, $Decay(L,k)$ denotes the ERR-based decay function defined over the ranked list, $Utility(L,k)$ denotes the utility of the (Search Engine Result Page) for the group of users who abandon the SERP at rank k, and $DistrSim^m(L,k)$ is the similarity between the distribution over groups (in the m-th attribute set) achieved by the top k documents in the SERP and the *target distribution* for the same attribute set. The weights w_i ($i = 0, \dots, M$) balance the *Utility* and *DistrSim*'s. At the NTCIR-17 FairWeb-1 task, movie topics (M-topics) had $M = 2$ attribute sets: RATINGS (number of ratings in IMDb) and ORIGIN (geographical region based on "country of origin" in IMDb); researcher topics (R-topics) also had $M = 2$ attribute sets: HINDEX (Google Scholar h-index) and GENDER (whether "he" or "she" was used in the researcher biography, or not); YouTube topics (Y-topics) had $M = 1$ attribute set: SUBSCS (number of subscribers of the content uploader). Participating groups were encouraged to develop systems that provide more exposure to entities (i.e., movies, researchers, or YouTube videos) that have not received much of it *and* achieve high SERP quality in terms of relevance.

Sakai et al. [25] explain that GFR is a measure of "expected user experience" over a user population for a given information need,[4] and discuss its advantages over a measure used at the TREC 2022 Fair Ranking Track [8], which has been discontinued. One important advantage of GFR is that it compares distributions over ordinal groups (such as the RATING, HINDEX and SUBSCS groups of the FairWeb-1 task) using appropriate *divergences*: more specifically, while GFR uses Jensen-Shannon Divergence (JSD) for nominal groups (such as ORIGIN and GENDER groups of the FairWeb-1 task), it uses Normalised Match Distance (NMD) and Root Normalised Order-aware Divergence (RNOD) for ordinal groups [20].

2.2 S-measure and M-measure

S-measure [23], which was adopted at the NTCIR 1CLICK tasks [24], is a measure for evaluating a textual summary returned in response to a query. One novel feature of S-measure is that it discounts the values of *iUnits* (information units or *nuggets*, which represent atomic pieces of relevant information) within the summary based on the nugget position, just like nDCG discounts the value of each

[4]This generalises the *Normalised User Utility* of Sakai and Robertson [22].

document based on its rank. Let V be the set of iUnits found within the summary, and for each iUnit $v \in V$, let $pos(v)$ denote the position of v within the summary. Moreover, let $w(v)$ be the weight of v, which reflects its relevance grade. Then S-measure can be expressed as follows.

$$S\text{-measure} = \frac{1}{\mathcal{N}} \sum_{v \in V} w(v) \max(0, 1 - pos(v)/L) \ . \quad (2)$$

Here, $\mathcal{N}$ is a normalisation factor, and L is a parameter that reflects user patience: at the 1CLICK tasks where Japanese summaries were evaluated, the organisers let $L = 1,000$ based on the observation that the average reading speed of Japanese text is around 500 characters per minute, and the assumption that the user needs to gather information within two minutes. Thus, the value of an iUnit decreases linearly as $pos(v)$ becomes larger, and becomes nought when $pos(v) > 1000$.

NTCIR-11 introduced the MobileClick task [13] as a successor of the 1CLICK task: the new task attempted to evaluate two-layered summaries, where the first-layer summaries contained several clickable links that represented possible intents for a given query. The intent probabilities were regarded as the transition probabilities from the first-layer summary to one of the second-layer summaries; the resultant extension of S-measure was named *M-measure.* While the two-layer summary framework resembles a two-round conversation that branches out depending on the previous turn, these previous NTCIR tasks considered the *relevance* of information only, unlike the present study.

2.3 Other Related Work

The TREC 2022 Conversational Assistance Track [17] evaluated multi-turn conversational search using a topic tree structure; while the conversations are evaluated in terms of relevance, conciseness, and naturalness, they did not consider group fairness. The track has been discontinued.

Inspired by the aforementioned S-measure, our proposed framework uses *word counts* as the basis for discounting a value of a relevant piece of information (i.e., a nugget). In the instantiation of our framework that is discussed in the present study, *all* words from a conversation (including those from user turns) contribute to the word count. This means that, if the user enters a long query, that is considered a large *cost*, and the values of systems' nuggets decrease accordingly. This mechanism is related to the discussion of *typing cost* or *cost of querying* in the economic interaction models of Azzopardi and Zuccon [3]. Earlier studies that consider the cost of querying (in terms of time spent) include Baskaya et al. [4] and Azzopardi et al. [2].

Regarding the evaluation of LLM-based systems, one obvious approach would be to build an LLM-based evaluator. However, if a black box LLM is evaluated with another black box LLM, it is our opinion that we may not be able to learn much: the two may be based on similar training data with similar algorithms and therefore the evaluator may overrate the system being evaluated; moreover, it may be difficult for an end-to-end LLM-based evaluator to explain to us why the system is good/bad and where exactly the problem lies. See also Faggioli et al. [9] and a recent Dagstuhl Seminar report [5, p.47] for related discussions. We argue that evaluation methods should be as simple and explainable to the researchers as possible, and our proposed method is based on this view.

3 PROPOSED EVALUATION METHOD

3.1 GFRC: A Generic Formulation

Our proposal is to combine the ideas of GFR and S-measure in order to evaluate a series of textual system turns based on both relevance and group fairness. Suppose that we want to evaluate a T-round user-system textual conversation $C = (U_1, S_1, \ldots, U_T, S_T)$. Let n_{ij} denote the j-th *nugget* in the i-th system turn S_i; more specifically, we define n_{ij} to be any substring of S_i that represents a *relevant* piece of information in the context of the previous turn sequence $(U_1, S_1, \ldots, U_i)$ as well as the previous nuggets in S_i, i.e., $(n_{i1}, \ldots, n_{i(j-1)})$. (For convenience, hereafter we denote n_{ij} as an *element* of *set* S_i.) Furthermore, let $pw(n)(\in [0, 1])$ denote the position-based weight of nugget n, where the position is defined in the context of conversation C. (We shall introduce an instantiation of $pw(n)$ in Section 3.2.) Let $g(n)(\in [0, 1])$ denote the relevance score (or gain value) of nugget n. We define the relevance-based score of conversation C as:

$$R(C) = \frac{1}{\mathcal{N}} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} pw(n_{ij}) g(n_{ij}) \ , \quad (3)$$

where $\mathcal{N}$ is a normalisation factor.

Next, we describe how the same conversation C can be evaluated in terms of group fairness, given M attribute sets and a target distribution (a probability mass function, to be more specific) D^m_* for each $m(= 1, \ldots, M)$. For every system turn S that contains a relevant nugget, we first compute an *achieved distribution* $D^m(S)$ using one of the following two possible methods. The *independent distribution* method computes $D^m(S)$ based solely on the group memberships of relevant nuggets contained in S; the *cumulative distribution* method computes $D^m(S)$ based not only on relevant nuggets in S but also on relevant nuggets observed in the *previous* system turns.[5] On the other hand, for any system turn that does not contain a relevant nugget, we can either ignore it, or treat it as if its achieved distribution is uniform (because if the turn does not mention any relevant entity, it is not introducing any bias towards any entity group and therefore "fair").

Let $PW(S)$ denote the position-based weight of *turn* S; we shall instantiate it later. The Group Fairness (GF) score of conversation C can be computed as:

$$GF(C) = \frac{1}{\mathcal{N}'} \sum_{i=1}^{T} PW(S_i) \sum_{m=1}^{M} w_m DistrSim^m(D^m(S_i) \parallel D^m_*) \ , \quad (4)$$

where $\mathcal{N}'$ is a normalisation factor and w_m is a weight assigned to the m-th attribute set s.t. $\sum_{m=1}^{M} w_m = 1$. Following GFR, we use JSD for computing the *DistrSim* function if the attribute set contains *nominal* groups, and we use either NMD or RNOD if the attribute

[5]The latter approach resembles the GFR framework [25], which computes an achieved distribution for each *user group* that is assumed to abandon the ranked list at a particular relevant document at rank r; the group memberships of all relevant entities in the top r documents contribute to the achieved distribution, which means that the group memberships of relevant entities near the top ranks contribute relatively heavily to the overall GFR score.

set contains *ordinal* groups [25]. Note that the *DistrSim* value is obtained as *one minus a divergence* such as JSD.

As a "quick summary" measure for ranking systems, the following combined measure may be of some use:

$$GFR(C) = \alpha R(c) + (1-\alpha) GF(C) \,, \tag{5}$$

where the α is a parameter that balances relevance and group fairness. In practice, however, we recommend to report R and GF scores separately and to visualise the relationship between the two.

3.2 GFRC: An Instantiation

In this section, we discuss a specific and practical instantiation of GFRC.

First, we provide a specific definition of $pw(n)$, the position-based nugget weight. Unlike the NTCIR 1CLICK tasks which primarily dealt with Japanese summaries, we discuss evaluating English textual conversations, and therefore consider word count-based positions rather than character count-based ones. Let us assume that the user is willing to spend up to 5 minutes to satisfy a particular information need; furthermore, we assume that the average reading speed of the user is 250 words per minute.[6] It then follows that the user is willing to read up to $L = 1,250$ words. For any word w from conversation C, let $wc(w)$ denote its *word count*: for example, for the first word in user turn U_1, $wc = 1$. Furthermore, for any relevant nugget n within a system turn, we define its word count $WC(n)$ as the word count of the *last* word that n corresponds to. Then we instantiate the position-based nugget weight in Eq. 3 as:

$$pw(n) = \max(0, 1 - \frac{WC(n)-1}{L}) \,, \tag{6}$$

so that the nugget weight linearly decreases as the conversation proceeds, and any nugget beyond the word count limit of 1,250 will be ignored. The accompanying normalisation factor $\mathcal{N}$ (See Eq. 3) can then be given by:

$$\mathcal{N} = \sum_{l=1}^{L} \left(1 - \frac{l-1}{L}\right) = \frac{L+1}{2} \,. \tag{7}$$

This is a "hard" normalisation factor, which represents a practically unattainable situation where every word in conversation C represents a relevant nugget. Nevertheless, this can be applied if we do not want the score to exceed 1.[7]

S-measure, which was designed for evaluating textual summaries, used a "softer" normalisation factor based on a "minimal" summary (*pseudo minimal output* [23]), which is analogous to the ideal ranked list of nDCG. However, this approach assumes a recall base of nuggets, and is probably not suitable for conversational search where it is often not possible to enumerate all relevant nuggets for evaluation.

Next, we instantiate the position-based weight for *turn* S to implement Eq. 4. While it is possible to make the weight *actually* position-aware by letting $PW(S)$ a function of $pw(n) (n \in S)$, e.g., the maximum nugget weight within S, in the present study, we consider a simpler option: let $\mathcal{S}$ be the set of system turns *excluding* those that do not contain any relevant entity, and simply let $PW(S_i) = 1$ iff $S_i \in \mathcal{S}$, with $\mathcal{N}' = |\mathcal{S}|$. That is, we simply average the *DistrSim*'s over relevant system turns. Similarly, when considering M different attribute sets in Eq. 4, we simply average the *DistrSim*'s across them.

In summary, our instantiations of $R(C)$ (Eq. 3) and $GF(C)$ (Eq. 4) are:

$$R(C) = \frac{2}{L+1} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} \max(0, 1 - \frac{WC(n_{ij})}{L}) g(n_{ij}) \,, \tag{8}$$

$$GF(C) = \frac{1}{m} \sum_{m=1}^{M} GF^m(C) \,, \tag{9}$$

where

$$GF^m(C) = \frac{1}{|\mathcal{S}|} \sum_{S_i \in \mathcal{S}} DistrSim^m(D^m(S_i) \parallel D^m_*) \,. \tag{10}$$

or $GF(C) = 0$ if $\mathcal{S} = \phi$ (i.e., none of the system turns are relevant).

4 A PILOT EXPERIMENT

4.1 Task and Topics

To illustrate how GFRC scores can be computed from user-system conversations, we conducted a pilot experiment using the entire 45 topics from the NTCIR-17 FairWeb-1 task, which we downloaded from the aforementioned FairWeb-1 website. The test topic set contains 15 M-topics, 15 R-topics, and 15 Y-topics.

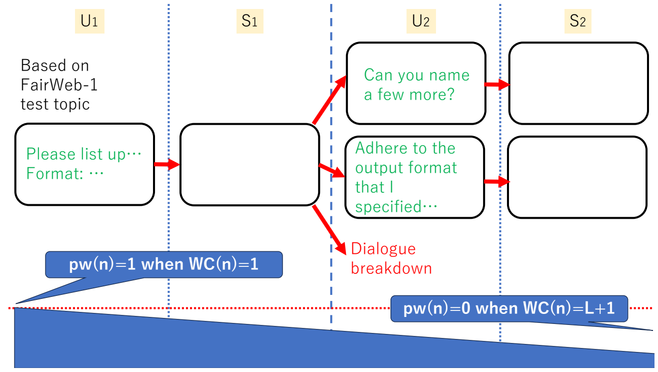

Figure 1: A two-round conversation protocol used in the pilot experiment.

Figure 1 shows the two-round ($T = 2$) conversation protocol we used to harvest textual interactions with a given conversational search system. (The bottom part of the figure shows a conceptual diagram of the linearly decaying position-based nugget weights.) The author of this paper manually interacted with conversational search systems on a web browser and recorded all textual exchanges. User Turn 1 (U_1) asks the system to list entities according to a FairWeb-1 topic. For example, U_1 for Topic R001 was:

```
Please list up researchers who have published at least
 one paper (any track) at the CHIIR conference.
Format: researcher name, URL
```

[6]The New Bing said (as of September 7, 2023): "According to a meta-analysis of 190 studies on reading speed, the average silent reading speed for an adult person is 238 words per minute (WPM) for non-fiction, while the average reading speed for fiction is 260 WPM"; Google Bard returned the same WPM numbers (same day); we have checked multiple websites to see that these answers are not hallucinations.

[7]An alternative would be score *standardisation* [19, 27, 28].

This is a form of *zero-shot prompting*.[8] As the figure indicates, System Turn 1 (S_1) may cause a *dialogue breakdown* [11]: the user is unable to continue the conversation due to highly inadequate responses such as:

```
I'm a text-based AI, and that is outside of my capabilities.
```

Otherwise, the same author entered U_2, which was either

```
Can you name a few more?
```

or one of the following depending on the topic type, as systems often fail to follow the output format specified in U_1:

```
Adhere to the output format I specified:
 "Format: movie title, IMDb URL"
Adhere to the output format I specified:
 "Format: researcher name, URL"
Adhere to the output format I specified:
 "Format: video title, youtube URL"
```

The entire conversations that we harvested in this way can be found in our supplementary material package.[9]

4.2 Systems

On September 2, 2023, the author of this paper followed the above protocol and interacted with both the New Bing and Google Bard. For each topic type, the topics were tested in the original order, and both systems were tested in parallel. For example, Topic M001 was tested with the New Bing and then with Google Bard; then the author moved on to Topic M002, and so on. Moreover, as LLM-based systems behave nondeterministically, after completing the conversations for all 45 topics, we repeated the same experiment. Thus, we had two trials with the New Bing (denoted by B1 and B2), and two trials with the Google Bard (denoted by G1 and G2). Note that we tested the topics in the same order in the two trials, as we wanted to avoid confounding the effect of topic ordering with that of nondeterministic responses. As we shall report in Section 5, we have evidence that a previous topic can actually affect the system response for the current topic.

5 SYSTEM RESPONSE OVERVIEW

Figures 2-4 visualise the outcomes of the two-round conversations with the New Bing and Google Bard using the M-topics, R-topics, and Y-topics from FairWeb-1. In each cell, a "o" means that the system turn "looks" useful to some extent; that is, it provides some URLs as requested (although they may not necessarily be relevant as we shall discuss later). A "x" means that the system turn is not useful: it either does not contain useful URLs (for example, for M001, B1 S_1 lists movies without providing IMDb URLs), or it refuses the user request. For example, for M001, G1 S_1 says:

```
I'm unable to help you with that, as I'm only a language model
 and don't have the necessary information or abilities.
```

A "-" indicates a dialogue breakdown (for example, for M001, G1 S_2 is a "-" due to S_1 shown above.) As the three figures show, all of the following patterns were observed in the experiment:

"o o": (shown in green) Both turns returned seemingly relevant URLs;

"o x": (shown in light green) In response to U_2: "`Can you name a few more?`" the system failed to respond adequately.

"x o": (shown in yellow) In response to U_2: "`Adhere to the output format that I specified...`" the system managed to improve the response format.

"x x": (shown in orange) In response to U_2: "`Adhere to the output format that I specified...`" the system still failed.

"x -": (shown in grey) S_1 caused a dialogue breakdown (e.g., the aforementioned case with G1 for M001).

At the bottom of each figure, some simple statistics are shown; they are not the focus of this paper. In each column, "%useful" is the proportion of "o" responses; In the columns for Turn 2, we have "ave%useful" which simply averages the "%useful" over the two turns; we also have "%recovery," defined as: "Of all topics for which the system's first turn was unsuccessful, what is the proportion that the system managed to return a useful second turn?" For example, in Figure 2, G2 has 14 unsuccessful S_1's, but recovered with S_2 for 9 of the topics, and therefore %recovery= 9/14 = 64.3%.

From Figures 2-4, the following observations can be made.

- First and foremost, the behaviour of each system is vastly different across the two trials, even though they were tested on the same day. Moreover, once they start to perform inadequately, they tend to continue to do so. For example, compare B1 and B2 in Figure 2 (M-topics), and G1 and G2 in Figure 3 (R-topics); in both cases, the second trial was a disaster, even though the first was not as poor. This extremely unstable nature of LLM-based conversational search systems poses a substantial challenge in terms of evaluation: the results of a small-scale experiment such as the one reported in the present study (let alone a few cherry-picked anecdotes we sometimes see in recent LLM papers!) are far too unreliable and unlikely to be generalisable.
- LLMs are known to hallucinate and, not surprisingly, the New Bing and Google Bard are no exceptions. For example, as indicated in Figure 3 footnotes 1 and 3, both of these systems hallucinated about the AIRS (Asia Information Retrieval Societies) conference: in fact, the final AIRS conference (not directly related to SIGIR) took place in 2019,[10], and instead the SIGIR-AP (Asia-Pacific) conference was launched in 2023.[11] Moreover, the URLs returned by Google Bard were often incorrect: for example, as we shall see later, most of the IMDb URLs that it returned for Topic M002 (time travel movies) were those for wrong movies.
- Google Bard behaved in the "o x" (light green) pattern very often (for R-topics and Y-topics), while this never happened with Bing in our experiment. Google's S_2 often did not make sense: even though S_1 contained some seemingly relevant URLs and therefore the system seems capable of performing the task to some extent, when asked for "a few more," S_2 was often "`I'm unable to help you with that, as I'm only a language model and don't have the necessary information or abilities.`" or something similar.
- Interestingly, during the conversation for R012 (about SIGIR) with Bing (Trial 2), in response to "`Adhere to the output format I specified: "Format: researcher`

[8]https://www.promptingguide.ai/
[9]http://waseda.box.com/evia2023pack
[10]https://link.springer.com/book/10.1007/978-3-030-42835-8
[11]http://www.sigir-ap.org/sigir-ap-2023/

	B1		B2		G1		G2	
	S1	S2	S1	S2	S1	S2	S1	S2
M001	x	o	x	x	x	-	x	x
M002	o	o	x	x	x	o	x	o
M003	o	o	x	x	x	o	x	o
M004	o	o	x	x	x	o	x	o
M005	o	o	x	x	x	o	o	o
M006	x	o	x	x	x	x	x	x
M007	x	o	x	x	x	o	x	o
M008	x	x	x	x	x	x	x	o
M009	x	x	x	x	x	x	x	o
M010	x	x	x	-	x	x	x	x
M011	x	x	x	x	x	o	x	o
M012	x	x	x	x	x	-	x	-
M013	x	x	x	x	x	o	x	o
M014	x	x	x	x	x	x	x	o
M015	x	x	x	x	x	x	x	x
%useful	26.7	46.7	0.0	0.0	0.0	46.7	6.7	66.7
ave%useful		36.7		0.0		23.3		36.7
%recovery		27.3		0.0		46.7		64.3

Figure 2: Overview of system responses for the M topics. B1, G1 etc. mean "Bing Trial 1" and "Google Trial1" and so on. "o" means that the system turn "looks" useful; "x" means that it is not useful; "-" means that there was no second system turn due to dialogue breakdown.

	B1		B2		G1		G2	
	S1	S2	S1	S2	S1	S2	S1	S2
R001	x	-	x	-	x	o	x	-
R002	x	-	x	-	o	x	x	-
R003	x	-	x	-	o	x	x	x
R004	x	-	x (*1)	-	x (*3)	-	x	-
R005	x	-	x	-	o	x	x	-
R006	x	-	x	x	o	x	x	-
R007	x	-	x	x	o	o (*4)	x	-
R008	x	-	x	-	o	o	x	-
R009	x	-	x	x	o	o	o	x
R010	x	-	x	x	o	x	x	-
R011	x	-	x	-	o	o	x	-
R012	x	-	x	x (*2)	o	o	x	-
R013	x	-	x	x	o	x	o	o
R014	x	-	x	x	o	x	x	-
R015	x	-	x	x	o	x	x	-
%useful	0.0	0.0	0.0	0.0	86.7	33.3	13.3	6.7
ave%useful		0.0		0.0		60.0		10.0
%recovery		0.0		0.0		100.0		0.0

Figure 3: Overview of system responses for the R topics. B1, G1 etc. mean "Bing Trial 1" and "Google Trial1" and so on. "o" means that the system turn "looks" useful; "x" means that it is not useful; "-" means that there was no second system turn due to dialogue breakdown.

*1: contains an obvious hallucination which caused a dialogue breakdown: "I found the proceedings of the AIRS 2022 conference"
*2: confusion with the previous topic "Here are some researchers who have published at least one paper (any track) at the KDD conference"
*3: contains an obvious hallucination which caused a dialogue breakdown: "the AIRS conference has been renamed to the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) since 2019"
*4: the 2nd list has an overlap with the 1st list

`name, URL`"", the system responded with "`Here are some researchers who have published at least one paper (any track) at the KDD conference...`" (Figure 3 footnote 2). This is because the previous topic discussed was R011 (about KDD). Hence previous conversations do seem to affect the current response in some cases, despite the fact that the systems kept ignoring the same output format instructions in many conversations.[12] Because the system responses are dependent on previous context, this adds a further challenge to the evaluation of conversational search: the sampled system turns are clearly not independent of one another.

[12] As of September 2, 2023, the New Bing forced the user to "start a new topic" after 30 system turns.

	B1		B2		G1		G2	
	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
Y001	x	o	x	o	x	-	x	-
Y002	o	o (*1)	o	o (*1)	o	x	o	x
Y003	o	o (*1)	x	o	x	-	o	N/A (*7)
Y004	x (*2)	-	o	o	o	x	o	x
Y005	o	o	x	x	x	-	x	-
Y006	o	o	x	-	x	-	x	-
Y007	o (*3)	o (*1)	x	o	o	o	o	x
Y008	o	o (*1)	x	o	o	x	o	x
Y009	o	o (*1)	x	o	o	x	o	x
Y010	o	o (*4)	x	o	x	-	x	-
Y011	o	o	x	-	x	-	x	x
Y012	o	o	x	o	o	x	o	N/A (*7)
Y013	o (*4)	o (*5)	x	o (*6)	x	-	x	-
Y014	o	o	x	-	o	x	x	-
Y015	o (*3)	o (*4)	x	x	x	-	x	-
%useful	66.7	40.0	13.3	53.3	46.7	6.7	46.7	0.0
ave%useful		53.3		33.3		26.7		23.3
%recovery		100.0		53.8		0.0		0.0

Figure 4: Overview of system responses for the Y topics. B1, G1 etc. mean "Bing Trial 1" and "Google Trial1" and so on. "o" means that the system turn "looks" useful; "x" means that it is not useful; "-" means that there was no second system turn due to dialogue breakdown.

*1: the 2nd list has an overlap with the 1st list
*2: system suddenly started apologising in Japanese
*3: contains duplicate URLs with different titles
*4 the same URL is repeated 5 times with different titles
*5 same as *4, and the URL is the same as the one in Turn 1
*6 the same URL is repeated 3 times with different titles
*7 experimenter failed to record the response, but probably it was "I'm just a language model" etc., i.e., an "x."

- Even when both S_1 and S_2 looked useful, for R-topics and Y-topics (especially the latter), the URLs returned sometimes had an overlap across the two turns, and/or had duplicate URLs within a single list. We argue that such duplicates should not be rewarded when evaluating the system turns: it seems sensible to treat all duplicate entities as nonrelevant, as was proposed at NTCIR-4 in the context of factoid question answering evaluation where only one answer string from each *equivalence class* of relevant answers was considered relevant in a ranked answer list [18].

The discussion above demonstrates that current LLM-based conversational search engines have a lot of room for improvement, and also that evaluating them in a reliable manner is highly challenging. We leave the grand challenge for future work; nevertheless, as one small piece that will hopefully contribute to the above challenge, we henceforth discuss a case study in which we apply our proposed evaluation framework to a topic from the above experiment.

6 A CASE STUDY

This section demonstrates how our measures can be computed from real conversational search data. As an example, we chose the Trial 1 results of B1 and G1 for Topic M002 "time travel movies" (highlighted in red in Figure 2) for the following reasons:

- M-topics are the only ones with two attribute sets where one of them allows *soft group membership*: one movie may belong to multiple ORIGIN groups (i.e., geographic regions) if its *country of origin* field on the IMDB page mentions multiple countries:[13] hence we can demonstrate how group fairness can be quantified in such evaluation settings;
- For M002, both the New Bing and Google Bard managed to return results that are not useless;
- The author of this paper was the Gold annotator of this FairWeb-1 topic; hence the author gets to decide which movies are relevant or not to his own information need.

6.1 Gold Annotation

The present author examined two text files that contained all the textual interactions between himself and the system (either the New Bing or Google Bard, Trial 1). On September 5, 2023 (i.e., 3 days after we harvested the conversations), the author examined the texts and annotated them by inserting lines after each entity returned by the system. Figure 5 shows the results: the annotations are shown in red. For example, in the New Bing's Turn 1 (S_1), the IMDb URL for "Back to the Future" is followed by the following pieces of information:

- This entity is L2-relevant (i.e., highly relevant) to the information need;
- On the IMDb page, the movie currently has 1.3M ratings and therefore falls into Group 4 (i.e., movies with 1M user ratings or more) of the RATINGS attribute set;

[13]The mapping from country names to eight *geographic regions* is also documented on http://sakailab.com/fairweb1/.

```
##### M002
### U1
Please list up sci-fi movies that feature time travel, listed in IMDb.
Format: movie title, iMDB URL
### S1 o
Here are some sci-fi movies that feature time travel, listed in IMDb:
Back to the Future (1985): https://www.imdb.com/title/tt0088763/
#L2
#RATINGS: (0,0,0,1) #1.3M
#ORIGIN: (0,1,0,0,0,0,0,0) #United States
wc= 35
The Terminator (1984): https://www.imdb.com/title/tt0088247/
#L2
#RATINGS: (0,0,1,0) #898K
#ORIGIN: (0,1/2,0,0,0,1/2,0,0) #United Kingdom,United States
wc= 39
Groundhog Day (1993): https://www.imdb.com/title/tt0107048/
#L2
#RATINGS: (0,0,1,0) #661K
#ORIGIN: (0,1,0,0,0,0,0,0) #United States
wc= 43
Interstellar (2014): https://www.imdb.com/title/tt0816692/
#L1
#RATINGS: (0,0,0,1) #2M
#ORIGIN: (0,2/3,0,0,0,1/3,0,0) #United States,United Kingdom,Canada
wc= 46
The Time Machine (1960): https://www.imdb.com/title/tt0054387/
#L2
#RATINGS: (0,0,1,0) #43K
#ORIGIN: (0,1,0,0,0,0,0,0) #United States
wc= 51
I hope this helps! Let me know if there's anything else I can help you with.
### U2
Can you name a few more?
### S2 o
Sure! Here are some more sci-fi movies that feature time travel, listed in IMDb:
12 Monkeys (1995): https://www.imdb.com/title/tt0114746/
#L2
#RATINGS: (0,0,1,0) #634K
#ORIGIN: (0,1,0,0,0,0,0,0) #United States
wc= 91
The Butterfly Effect (2004): https://www.imdb.com/title/tt0289879/
#L2
#RATINGS: (0,0,1,0) #508K
#ORIGIN: (0,1,0,0,0,0,0,0) #United States,Canada
wc= 96
Looper (2012): https://www.imdb.com/title/tt1276104/
#L2
#RATINGS: (0,0,1,0) #592K
#ORIGIN: (0,1/2,0,1/2,0,0,0,0) #United States,China
wc= 99
Edge of Tomorrow (2014): https://www.imdb.com/title/tt1631867/
#L2
#RATINGS: (0,0,1,0) #711K
#ORIGIN: (0,1,0,0,0,0,0,0) #United States,Canada
wc= 104
Predestination (2014): https://www.imdb.com/title/tt2397535/
#L2
#RATINGS: (0,0,1,0) #294K
#ORIGIN: (0,1/2,0,0,0,0,0,1/2)#Australia,United States
wc= 107
I hope this helps! Let me know if there's anything else I can help you with.
```

```
##### M002
### U1
Please list up sci-fi movies that feature time travel, listed in IMDb.
Format: movie title, iMDB URL
### S1 x
[A 452-word response that does not follow the above format]
### U2
Adhere to the output format I specified: "Format: movie title, iMDB URL"
### S2 o
Sure, here are the sci-fi movies that feature time travel, listed in IMDb, in the specified format:
Movie title    IMDb URL
Back to the Future (1985)  https://www.imdb.com/title/tt0088763/
#L2
#RATINGS: (0,0,0,1) #1.3M
#ORIGIN: (0,1,0,0,0,0,0,0) #United States
wc= 506
Back to the Future (1985) movie posterOpens in a new window
www.imdb.com
Back to the Future (1985) movie poster
Pulp Fiction (1994) https://www.imdb.com/title/tt0110912/
#L0
Pulp Fiction (1994) movie posterOpens in a new window
www.originalfilmart.com
Pulp Fiction (1994) movie poster
Looper (2012)       https://www.imdb.com/title/tt1853739/
#L0 # wrong URL "You're Next"
Looper (2012) movie posterOpens in a new window
www.imdb.com
Looper (2012) movie poster
Interstellar (2014)    https://www.imdb.com/title/tt0816692/
#L1
#RATINGS: (0,0,0,1) #2M
#ORIGIN: (0,2/3,0,0,0,1/3,0,0) #United States,United Kingdom,Canada
wc= 560
Interstellar (2014) movie posterOpens in a new window
www.imdb.com
Interstellar (2014) movie poster
Tenet (2020) https://www.imdb.com/title/tt10375988/
#L0 # wrong URL "The Eze-Nwanyi Kate Ajike Love Show"
Tenet (2020) movie posterOpens in a new window
www.imdb.com
Tenet (2020) movie poster
Doctor Strange (2016)      https://www.imdb.com/title/tt3498496/
#L0 # wrong URL "Michigan Pheasant"
Doctor Strange (2016) movie posterOpens in a new window
www.imdb.com
Doctor Strange (2016) movie poster
The Terminator (1984)      https://www.imdb.com/title/tt0089531/
#L0 # wrong URL "Madame Deborah"
Terminator (1984) movie posterOpens in a new window
www.imdb.com
Terminator (1984) movie poster
The Butterfly Effect (2004)        https://www.imdb.com/title/tt0322997/
#L0 # wrong URL "Cruiskeen Lawn"
Butterfly Effect (2004) movie posterOpens in a new window
www.imdb.com
Butterfly Effect (2004) movie poster
Primer (2004)       https://www.imdb.com/title/tt0389895/
#L0 # wrong URL "Crystal Force 2: Dark Angel"
Primer (2004) movie posterOpens in a new window
www.imdb.com
Primer (2004) movie poster
12 Monkeys (1995)  https://www.imdb.com/title/tt0112389/
#L0 wrong URL "The Thief and the Cobbler"
12 Monkeys (1995) movie posterOpens in a new window
www.imdb.com
12 Monkeys (1995) movie poster
I hope this is more helpful!
```

Figure 5: Bing Trial 1 (left) vs. Google Trial 1 (right) for Topic M002 (time travel movies) with relevance and group membership annotations, with word counts.

- On the IMDb page, United States is named as the *country of origin* of this movie; hence this movie belongs solely to Group 2 (America) of the ORIGIN attribute set.

In Figure 5, word count information is also provided in balloons. For example, the word count for the "Back to the Future" URL is 35: that is, this is the 35th word in this conversation. Recall that when we compute the word count, words from the user turns are *included* in the present study: that is, if the user enters many words, that is considered as a cost, and the system effectiveness is discounted accordingly.

6.2 Measure Computation

Table 1 shows how the R scores are computed for the conversations shown in Figure 5 based on Eq. 8; here, L2-relevant and L1-relevant entities are mapped to gain values of 1 and 0.5, respectively. It can be observed that while the New Bing managed to return 10 relevant entities (with "Interstellar" considered by the Gold annotator to be an L1-relevant entity unlike the other movies), Google Bard managed to return only 2: many of the IMDb URLs it returned were incorrect. Note that the rightmost column ($pw(n)g(n)$) is analogous to the discounted gain of nDCG: for example, while both Bing and Google return "Back to the Future" as the first L2-relevant entity in their responses, the $pw(n)g(n)$ for Google is lower (1.9456 for Bing but 1.1920 for Google). This is largely because, as Figure 5 (right) shows, Google's first turn did not contain any relevant nuggets *and wasted as many as 452 words.* As a result, the word count for the IMDb URL of the "Back to the Future" nugget is 506. According to Eq. 8 with $L = 1,250$, the R score for Bing is 0.0143 while that for Google is only 0.0014 (i.e., only about 9.8% of the R score for Bing).

Table 1: Computing the R scores for the conversations shown in Figure 5.

movie (Nugget n)	$pw(n)$	$g(n)$	$pw(n)g(n)$
(a) New Bing			
Back to the Future	0.9728	1	0.9728
The Terminator	0.9696	1	0.9696
Groundhog Day	0.9664	1	0.9664
Interstellar	0.9640	0.5	0.4820
The Time Machine	0.9600	1	0.9600
12 Monkeys	0.9280	1	0.9280
The Butterfly Effect	0.9240	1	0.9240
Looper	0.9216	1	0.9216
Edge of Tomorrow	0.9176	1	0.9176
Predestination	0.9152	1	0.9152
		R	0.0143
(b) Google Bard			
Back to the Future	0.5960	1	0.5960
Interstellar	0.5528	0.5	0.2764
		R	0.0014

Note also how the relevant entities from multiple turns are treated seamlessly in the R score computation: for example, Figure 5 (left) shows that 5 relevant entities were returned in System Turn 1 (S_1) and another 5 were returned in System Turn 2 (S_2); but from Table 1 it is clear that the R score treats all of these entities just as those that lie on top of a single slope (See Figure 1 bottom).

Table 2 shows how the GF scores are computed for the conversations shown in Figure 5 based on Eq. 10. First, let us discuss the left part of the table that computes GF scores for the RATINGS attribute set (with 4 ordinal groups, where Group 4 represents movies with over 1M ratings). From Figure 5 (left), the group membership vectors for the 5 relevant entities returned by Bing in S_1 are $(0, 0, 0, 1), (0, 0, 1, 0), (0, 0, 1, 0), (0, 0, 0, 1), (0, 0, 1.0)$. By averaging them we obtain the achieved distribution for S_1: $(0, 0, 0.6, 0.4)$ as shown in Table 2 (left). By comparing this achieved distribution with a uniform gold distribution over the 4 groups in terms of RNOD,[14] the *DistrSim* value is 0.6773 for S_1. Similarly, the achieved distribution of S_2 for Bing is $(0, 0, 1, 0)$ and the *DistrSim* in terms of RNOD is 0.4796. Finally, by applying Eq. 10, we obtain a RATINGS-based GF score of 0.5785 for Bing. On the other hand, as only S_2 contains relevant entities in the Google results, from Eq. 10, only the *DistrSim* value for S_2 contributes to the RATINGS-based GF score for Google, which amounts to 0.4049 (i.e., only about 70.0% of the score for Bing). It can be observed that Google underperforms Bing because it is more heavily biased towards "famous" movies (i.e., those with many ratings).

Finally, let us discuss the right part of Table 2 that computes GF scores for the ORIGIN attribute set (with 8 nominal groups representing geographic regions). Again, the achieved distributions for the system turns are computed by averaging the group membership vectors shown in Figure 5: for example, the distribution for Bing's S_1 $(0, 0.8333, 0, 0, 0, 0.1667, 0, 0)$ means that it is heavily biased towards *America* (mapped from United States and Canada), while acknowledging some presence of *Europe* (mapped from United Kingdom). As the groups (i.e., regions) are nominal this time, the *DistrSm* values are obtained using JSD. It can be observed that, even in terms of ORIGIN-based GF scores, Google slightly underperforms Bing: its score is about 95.8% of that of Bing (0.4303 vs 0.4493). This reflects the fact that Google only covers Groups 2 and 6 (America and Europe), while Bing covers Groups 4 and 8 (Asia and Oceania) in addition, as can be verified from Figure 5.

In summary, in this particular example, Bing outperforms Google in terms of all three measures: the R (relevance) score, the RATINGS-based GF score, and the ORIGIN-based GF score. This example suggests that our framework, which extends the GFR framework for evaluating ranked lists, may be useful for improving conversational search engine responses in terms of relevance and group fairness.

7 CONCLUSIONS

We have demonstrated how the GFR framework used in the NTCIR-17 FairWeb-1 task can be extended for evaluating a series of textual system turns in conversational search. By using the full test topic set of FairWeb-1 to harvest actual user-system conversations from the New Bing and Google Bard, we demonstrated how a series of system turns can be evaluated using our GFRC (Group Fairness and Relevance of Conversations) framework.

With the rapid progress of LLM-based conversational search systems and the various social problems that accompany it (e.g., hallucinations and biases), we argue that evaluating a sequence of system turns is becoming at least as important as evaluating a ranked list of documents, to put it mildly. Hence, the NTCIR FairWeb organisers are currently considering a new conversational subtask for NTCIR-18 along with the existing web page ranking task, so that group fairness can be studied by looking across these two paradigms. From the participants' point of view, can the same algorithm be used for group fair *ranking* and group fair *conversations*? From the organisers' point of view, how are GFR (for rankings) and GFRC (for conversations) related to each other, and how can conversation trees (i.e., turns that branch out) [17] be obtained so that *turn transition probabilities* can be incorporated into GFRC?

Our pilot experiment also exemplified some known challenges in evaluating conversational search. The following are a few open questions for human-in-the-loop conversational search evaluation that we have observed.

- Given the highly unstable and context-dependent nature of the system turns based on LLMs, how can we conduct reliable evaluation with appropriate sampling methods and sample sizes?
- How can we sample and evaluate *worst-case* situations (rather than *typical* ones represented by *mean* scores etc.), so that any possible harm on users can be detected in a timely manner?
- What is the appropriate methodology to combine human-in-the-loop evaluations (which we cannot do without, as satisfying and protecting the *users* is our goal) and *user simulations* [10, 15, 29] which enable large-scale evaluations and hence may enable a better coverage of the above-mentioned worst-case scenarios?

[14]NMD may be used instead as the groups are ordinal.

Table 2: Computing the GF scores for the conversations shown in Figure 5.

	$D^{\text{RATINGS}}(S_i)$				*DistrSim*	$D^{\text{ORIGIN}}(S_i)$								*DistrSim*
	1	2	3	4	(RNOD)	1	2	3	4	5	6	7	8	(JSD)
(a) New Bing														
S_1	0	0	0.6	0.4	0.6773	0	0.8333	0	0	0	0.1667	0	0	0.4303
S_2	0	0	1	0	0.4796	0	0.8	0	0.1	0	0	0	0.1	0.4682
GF^{RATINGS}					0.5785	GF^{ORIGIN}								0.4493
(b) Google Bard														
S_1	-	-	-	-		-	-	-	-	-	-	-	-	-
S_2	0	0	0	1	0.4049	0	0.8333	0	0	0	0.1667	0	0	0.4303
GF^{RATINGS}					0.4049	GF^{ORIGIN}								0.4303

REFERENCES

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861 [cs.CL]

[2] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. ACM, 23–32.

[3] Leif Azzopardi and Guido Zuccon. 2018. Economic Models of Interaction. In *Computational Interaction*, Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes (Eds.). Oxford University Press.

[4] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA). ACM, 105–114.

[5] Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr. 2023. Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. arXiv:2305.01509 [cs.IR]

[6] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of ACM CIKM 2009*. 621–630.

[7] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. arXiv:2105.05779 [cs.IR]

[8] Michael D Ekstrand, Graham McDonald, and Amifa Raj. 2022. Overview of the TREC 2021 Fair Ranking Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*. NIST.

[9] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan). ACM, 39–50.

[10] David Griol, Javier Carbó, and José M. Molina. 2013. AN AUTOMATIC DIALOG SIMULATION TECHNIQUE TO DEVELOP AND EVALUATE INTERACTIVE CONVERSATIONAL AGENTS. *Applied Artificial Intelligence* 27, 9 (2013), 759–780.

[11] Ryuichiro Higashinaka, Kotato Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož, Slovenia). European Language Resources Association (ELRA), 3146–3150.

[12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[13] Makoto P. Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. 2014. Overview of the NTCIR-11 MobileClick Task. In *Proceedings of NTCIR-11*. 195–207. https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-MOBILECLICK-KatoM.pdf

[14] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[15] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM TOIS* 39, 4, Article 51 (2021).

[16] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374 [cs.AI]

[17] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas4, and Svitlana Vakulenko. 2023. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *NIST Special Publication 500-338: The Thirty-First Text REtrieval Conference Proceedings (TREC 2022)* (Virtual Event). NIST.

[18] Tetsuya Sakai. 2004. New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering. In *Working Notes of NTCIR-4*. https://research.nii.ac.jp/ntcir/ntcir-ws4/NTCIR4-WN/OPEN/OPENSUB_Tetsuya_Sakai.pdf

[19] Tetsuya Sakai. 2016. A Simple and Effective Approach to Score Standardisation. In *Proceedings of ACM ICTIR 2016* (Newark, Delaware, USA). ACM, 95–104.

[20] Tetsuya Sakai. 2020. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online). Association for Computational Linguistics, 2759–2769. https://aclanthology.org/2021.acl-long.214.pdf

[21] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. arXiv:2305.08290 [cs.IR]

[22] Tetsuya Sakai and Noriko Kando. 2008. Modelling A User Population for Designing Information Retrieval Metrics. In *Proceedings of EVIA 2008*. 30–41. https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/07-EVIA2008-SakaiT.pdf

[23] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. 2011. Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access. In *Proceedings of CIKM 2011* (Glasgow, Scotland, UK). ACM, 621–630.

[24] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. 2011. Overview of NTCIR-9 1CLICK. In *Proceedings of NTCIR-9* (Tokyo, Japan). 180–201.

[25] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A Versatile Framework for Evaluating Ranked Lists in terms of Group Fairness and Relevance. *ACM TOIS* (2023). https://dl.acm.org/doi/pdf/10.1145/3589763

[26] Sijie Tao, Nuo Chen, Tetsuya Sakai, Zhumin Chu, Hiromi Arai, Ian Soboroff, Nicola Ferro, and Maria Maistro. 2023. Overview of the NTCIR-17 FairWeb-1 Task. In *Proceedings of NTCIR-17*. to appear. https://doi.org/10.20736/0002001318

[27] Julián Urbano, Harlley Lima, and Alan Hanjalic. 2019. A New Perspective on Score Standardization. In *Proceedings of ACM SIGIR 2019* (Paris, France). ACM, New York, NY, USA, 1061–1064.

[28] William Webber, Alistair Moffat, and Justin Zobel. 2008. Score Standardization for Inter-Collection Comparison of Retrieval Systems. In *Proceedings of ACM SIGIR 2008* (Singapore, Singapore). ACM, 51–58. https://doi.org/10.1145/1390334.1390346

[29] Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and Simulating User Utterance Reformulation in Conversational Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain). Association for Computing Machinery, 133–143.