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 GFR (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] 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 and Google Bard, we demonstrate how a series of system turns can be evaluated using our evaluation framework, which we call GFR (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.

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].

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

Here, \( \text{Decay}(L, k) \) denotes the ERR-based decay function defined over the ranked list, \( \text{Utility}(L, k) \) denotes the utility of the (Search Engine Result Page) for the group of users who abandon the SERP at rank k, and \( \text{DistrSim}^m(L, k) \) 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 \( w_i (i = 0, \ldots, M) \) 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, 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 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 nugget position, just like nDCG discounts the value of each.

\[\text{nDCG} = \sum_{k=1}^{1|L|} \frac{2^{r_k} - 1}{\log(k+1)} \]

This generalises the Normalised User Utility of Sakai and Robertson [22].
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 \( n_{ij} \) as any substring of \( S_i \) that represents a relevant piece of information in the context of the previous turn sequence \((U_1, S_1, \ldots, U_j)\) 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 \( S_i \).) 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:

\[
R(C) = \frac{1}{N} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} pw(n_{ij}) g(n_{ij}),
\]

where \( N \) 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) \( D^m_n \) 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:

\[
GF(C) = \frac{1}{N'} \sum_{i=1}^{T} PW(S_i) \sum_{n=1}^{M} w_m \text{DistrSim}^m(D^m(S_i) \mid D^m_n),
\]

where \( N' \) 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 GF score.
set contains ordinal groups [25]. Note that the DistriSim 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:

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

where the \(\alpha\) 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. 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:

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

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 \(N\) (See Eq. 3) can then be given by:

$$N = \sum_{i=1}^{L} \left(1 - \frac{i - 1}{L}\right) = \frac{L + 1}{2}.\quad \text{(7)}$$

This is a "hard" normalisation factor, which represents a practically unattainable situation where every 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 \(S\) be the set of system turns excluding those that do not contain any relevant entity, and simply let \(PW(S_i) = 1\) iff \(S_i \in S\), with \(N = |S|\). That is, we simply average the DistriSim’s over relevant system turns. Similarly, when considering \(M\) different attribute sets in Eq. 4, we simply average the DistriSim’s across them.

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

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

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

where

$$G_{P^m}(C) = \frac{1}{|S|} \sum_{D_i \in S} \text{DistriSim}^{P^m}(D^m_i) \parallel D^m_i.$$

or \(GR(C) = 0\) if \(S = \emptyset\) (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 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.
This is a form of zero-shot prompting. 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.

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 $U_2$: “Adhere to the output format that I specified…” the system managed to improve the response format.

"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, and instead the SIGIR-AP (Asia-Pacific) conference was launched in 2023. 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

---

Footnotes:

1. [https://www.prompthinguide.ai/](https://www.prompthinguide.ai/)
Figure 2: Overview of system responses for the M topics. B1, G1 etc. mean “Bing Trial 1” and “Google Trial 1” 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.

Figure 3: Overview of system responses for the R topics. B1, G1 etc. mean “Bing Trial 1” and “Google Trial 1” 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...” (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.
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; 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;

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

<table>
<thead>
<tr>
<th>movie (Nugget $n$)</th>
<th>$p_w(n)$</th>
<th>$g(n)$</th>
<th>$p_w(n)g(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) New Bing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back to the Future</td>
<td>0.9728</td>
<td>1</td>
<td>0.9728</td>
</tr>
<tr>
<td>The Terminator</td>
<td>0.9696</td>
<td>1</td>
<td>0.9696</td>
</tr>
<tr>
<td>Groundhog Day</td>
<td>0.9664</td>
<td>1</td>
<td>0.9664</td>
</tr>
<tr>
<td>Interstellar</td>
<td>0.9640</td>
<td>0.5</td>
<td>0.4820</td>
</tr>
<tr>
<td>The Time Machine</td>
<td>0.9600</td>
<td>1</td>
<td>0.9600</td>
</tr>
<tr>
<td>12 Monkeys</td>
<td>0.9280</td>
<td>1</td>
<td>0.9280</td>
</tr>
<tr>
<td>The Butterfly Effect</td>
<td>0.9240</td>
<td>1</td>
<td>0.9240</td>
</tr>
<tr>
<td>Looper</td>
<td>0.9216</td>
<td>1</td>
<td>0.9216</td>
</tr>
<tr>
<td>Edge of Tomorrow</td>
<td>0.9176</td>
<td>1</td>
<td>0.9176</td>
</tr>
<tr>
<td>Predestination</td>
<td>0.9152</td>
<td>1</td>
<td>0.9152</td>
</tr>
<tr>
<td>$R$</td>
<td></td>
<td></td>
<td>0.0143</td>
</tr>
</tbody>
</table>

| (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 $(0, 0, 0, 1), (0, 0, 1, 0), (0, 0, 1, 0), (0, 0, 0, 0), (0, 0, 1, 0)$. By averaging them we obtain the achieved distribution for $S_1$: $(0, 0, 0, 6, 0, 0, 4)$ as shown in Table 2 (left). By comparing this achieved distribution with a uniform gold distribution over the 4 groups in terms of RNODE, the $DistSim$ value is 0.6773 for $S_1$. Similarly, the achieved distribution of $S_2$ for Bing is $(0, 0, 1, 0)$ and the $DistSim$ in terms of RNODE 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 $DistSim$ 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 $S_1$ $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ 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 $DistSim$ 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 GFR (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?
Table 2: Computing the GF scores for the conversations shown in Figure 5.

<table>
<thead>
<tr>
<th>(a) New Bing</th>
<th>(b) Google Bard</th>
</tr>
</thead>
<tbody>
<tr>
<td>D^G RATING(S_i)</td>
<td>D^G RATING(S_i)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S_1</td>
<td>S_1</td>
</tr>
<tr>
<td>1 0 0.6 0.4</td>
<td>0 0 0.833 0.1</td>
</tr>
<tr>
<td>0.6773</td>
<td>0.4796</td>
</tr>
<tr>
<td>0.4049</td>
<td>0.4049</td>
</tr>
<tr>
<td>DistrSim(RNOD)</td>
<td>DistrSim(JSD)</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>1 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>0.1 0 0 0.1 0 0.1667 0 0</td>
<td>0.5785</td>
</tr>
<tr>
<td>0.8333 0.1 0 0.8333 0 0.1667 0 0</td>
<td>0.4049</td>
</tr>
</tbody>
</table>

REFERENCES


