Evaluating Parrots and Sociopathic Liars: A keynote at ICTIR 2023

Tetsuya Sakai
Waseda University
Japan
tetsuyasakai@acm.org

Abstract

At ICTIR 2023, which took place on July 23, 2023 in Taipei, I talked about the dark sides of the Large Language Model (LLM) era, and a tentative framework for auditing LLM-based conversational search systems to protect the users from undesirable system responses. This extended abstract provides a short summary of the one-hour keynote and provides pointers to related resources. In addition, it mentions a few recent developments relevant to the talk.

Date: 23 July 2023.

1 Introduction

By "parrots" I meant the *stochastics parrots* of Bender et al. [2021], that is, LLMs (in the context of conversational search). By "sociopathic liars," I meant the same thing, as Bowen provides the following definition.[1] "*Sociopathic liars are the most damaging types of liars because they lie on a routine basis without conscience and often without reason. Whereas pathetic liars lie to get along, and narcissistic liars prevaricate to cover their inaction, drama, or ineptitude, sociopaths lie simply because they feel like it. Lying is easy for them, and they lie with no conscience or remorse.*"

While many people (researchers, politicians, reporters, etc.) like to talk about how (some) people can benefit from LLM-based conversational search systems, I wanted to discuss with the audience the negative aspects of the advent of these systems, and in particular to talk about how we might be able to protect the users from undesirable system responses.

[1] https://sc.edu/study/colleges_schools/cic/journalism_and_mass_communications/news/2018/pr_prose_types_of_liars.php

2 A Short Summary

2.1 Dark Sides of the LLM Era

Clearly, I am not the first to point out these negative aspects of the LLM era, but I think we should keep talking about them and thinking about how to address them. Specifically, I first touched upon the following topics associated with the advent of LLM-based conversational search.

- Negative implications of society that relies on LLMs that often *hallucinate* (i.e., "lie") *with confidence*[2], and *flatter* users;[3]
- Social inequity, where some people enjoy the benefits of LLMs while others do not have the same privilege: for example, for users in some countries, the conversational search results may be less useful compared to elsewhere; they may even have more limited access to LLM-based APIs to begin with;
- Worker exploitation: labellers are hired for *alignment* purposes and are made to suffer;[4]
- LLMs giving out a lot of CO2 [Scells et al., 2022] and consuming a lot of water [Zuccon et al., 2023];
- Harms on science and research (cherry-picking, lack of transparency, repeatability, and reproducibility, *contamination* (i.e., evaluating with test data that may not be "clean"), prompting in the dark (i.e., lack of scientific explanation as to why and how certain prompts work while others do not).

Other problems include plagiarism and invasion on privacy (i.e., data theft issues), although they were not discussed explicitly in the keynote. I argued that researchers should not just ignore these clear and present problems.

2.2 Protecting Users from LLM-based Conversational Search Results

The main part of my keynote was about my SWAN (Stochastic Weighted Average Nugget) framework [Sakai, 2023c], which is a simple and generic score computation scheme for auditing/evaluating system responses in conversational search, where multiple evaluation criteria can be considered at the same time. The SWAN framework is expected to work as follows.

1. We (i.e., preferably people who do not have a COI with the system being evaluated) sample conversations through user experiments and/or user simulation. The former is necessary for collecting real conversations; the latter is necessary for protecting users from potential harm and for obtaining many possible conversations efficiently.
2. We employ LLMs to break the conversations into nuggets: Type-F ("factual") nuggets represent factual claims, while Type-O ("other") nuggets represent *dialogue acts* [Stolcke et al., 2000].

[2] *Calibration* refers to the ability of a system to align its confidence score with the accuracy of the associated answer [Liang et al., 2023]. In Table 1 (from Sakai [2023c]), this property is listed as *modesty*.

[3] Liu et al. [2023] discusses the *sycophancy* of LLM responses. In Table 1 (from Sakai [2023c]), the *sincerity* criterion is expected to penalise such behaviours.

[4] For example, see *Stochastic Parrots Day: On Worker Exploitation, Data Theft, and the Centralization of Power* (video): https://peertube.dair-institute.org/w/qBgQLX5DgMgHNF5NMo7886

3. We (semi-)automatically score system conversations at the nugget level (or at the turn-level) based on some criteria, possibly a subset of the taxonomy shown in Table 1.
4. Where necessary, we aggregate the scores across turns, across conversation sessions, and across criteria, while paying attention to individual phenomena (e.g., locating nuggets that receive low scores from the viewpoint of a particular criterion).

While Steps 2 and 3 will probably relie on LLMs, these subtasks are *compartmentalised*: each LLM does not know what the other LLMs are doing, and works on the given subtask to achieve high accuracy. This is different from evaluating LLMs with another black box LLM in an end-to-end manner: for related discussions, see Bauer et al. [2023]; Faggioli et al. [2023]; Thomas et al. [2023].

I also argued that conversational search auditing/evaluation schemes in general should satisfy the following requirements least:

Alertness Potential problems should not be missed at auditing time. Satisfying the "average" users is not enough; we need to detect potential harms on marginalised users.

Specificity We should be able to exactly locate the problem ("Where in which system turn is the problem?"); hence our use of nuggets as the default evaluation unit.

Versatility We should be able to handle task-oriented and non-task-oriented conversations seamlessly, to handle single-turn and multi-turn conversations seamlessly, and to consider multiple evaluation criteria.

Agility The auditing/evaluation procedure should keep up with the rapid progress of LLMs; hence the necessity of relying on LLMs to solve subtasks of LLM auditing/evaluation.

Transparency The process and results of auditing/evaluation should be easily interpretable.

Neutrality The auditing/evaluation should not favour any particular approach, and should not cherry-pick evaluation results.

The SWAN framework was designed from the above perspectives. The "S" (Stochastic) in "SWAN" implies that we will have to handle *trees* of conversations (user and system responses branching out as the conversation proceeds: see, for example, Owoicho et al. [2023]), although we have not yet tried sampling conversations in that way.

3 Relevant Links

Links related to the ICTIR 2023 keynote:

- One-page abstract in the ACM Digital Library ICTIR 2023 Proceedings[5]
- Slide deck (114 slides)[6]
- The SWAN arxiv paper (13 pages) [Sakai, 2023c][7]

[5] https://doi.org/10.1145/3578337.3605144
[6] https://waseda.box.com/ictir2023keynote-slides
[7] https://arxiv.org/abs/2305.08290

Table 1. 20(+1) criteria for evaluating textual conversational systems with SWAN (from Sakai [2023c]).

	Criterion	Brief comments (with related and (near-)equivalent criteria)
0	Fluency (solved)	(Naturalness) Does the turn pass as a manually composed text?
1	Coherence	(Relevance) Does the turn make sense as a response to the previous user turn?
2	Sensibleness	No common sense mistakes, no absurd responses
3	Correctness	Is the nugget factually correct?
4	Groundedness	Is the nugget based on some supporting evidence?
5	Explainability	Can the user see how the system came up with the nugget?
6	Sincerity	Is the nugget likely to be consistent with the system's internal results?
7	Sufficiency	(Recall) Does the turn satisfy the requests in the previous user turn?
8	Conciseness	Is the system turn minimal in length?
9	Modesty	(Confidence) Does the system's confidence about the nugget seem appropriate?
10	Engagingness	(Interestingness, Topic breadth) Does the system nugget/turn make the user want to continue the conversation?
11	Recoverability	Does the system turn keep the user interacting after the user has expressed dissatisfaction?
12	Originality	(Creativity) Is the nugget original, and not a copy of some existing text?
13	Fair exposure	Does the system mention different groups fairly across its turns?
14	Fair treatment	Does the system provide the same benefit to different users and user groups?
15	Harmlessness	(Safety, Appropriateness) No threats, no insults, no hate or harassment, etc.
16	Consistency	Given the nuggets seen so far, is the present nugget logically possible?
17	Retentiveness	Does the system "remember"?
18	Robustness to input variations	Does the system eventually provide the same information no matter how we ask?
19	Customisability	(Personalisability) Does the system adapt to different users and user groups?
20	Adaptability	Does the system keep up with the changes in the world?

Links related to the earlier ECIR 2023 online keynote (March 31, 2023), titled "On A Few Responsibilities of (IR) Researchers: Fairness, Awareness, and Sustainability."

- One-page abstract within the frontmatter of ECIR 2023 Proceedings (Volume I: LNCS 13980)[8]
- Slide deck (74 slides)[9]
- SIGIR Forum post-conference keynote extended abtract (June 23) [Sakai, 2023b][10]

4 Coming Up Next...

On December 13, 2023 at the NTCIR-17 Conference held in Tokyo, I will host a one-hour panel titled *Responsible Information Access: Fairness, Harmlessness, Sustainability, and More*, featur-

[8] https://link.springer.com/content/pdf/bfm:978-3-031-28244-7/1
[9] https://waseda.box.com/ecir2023keynote
[10] https://sigir.org/wp-content/uploads/2023/07/p04.pdf

ing Haruka Maeda (Kyoto University), Paul Thomas (Microsoft), and Mark Sanderson (RMIT University) as panelists. The plan for the panel in slide deck form is here[11].

Regarding group fairness evaluation for conversational search (Criterion 13 in Table 1), myself and colleagues are planning to propose the second FairWeb task for NTCIR-18 that features a new conversational search subtask. My EVIA 2023 paper (to be presented on December 12, 2023 at NTCIR-17) [Sakai, 2023a] discusses how LLM-based conversational search systems can be evaluated by extending the Group Fairness and Relevance (GFR) framework [Sakai et al., 2023] that was designed for evaluating ranked lists and used at the NTCIR-17 FairWeb-1 task.[12] The new measure, called GFRC (GFR for Conversation), can be interpreted as an instantiation of SWAN. We also plan to experiment with other instantiations of SWAN using different criteria from Table 1 in our future work.

Acknowledgements

I thank the ICTIR 2023 general chair (Masaharu Yoshioka) and PC chairs (Julia Kiseleva and Mohammad Aliannejadi) for giving me the opportunity for the keynote, and those who attended the keynote session and participated in the discussion. I also thank the SIGIR 2023 general chairs (Hsin-Hsi Chen, Wei-Jou Duh, and Hen-Hsen Huang) for supporting the ICTIR 2023 conference.

References

Christine Bauer, Ben Carterette, Nicola Ferro, Norbert Fuhr, and Guglielmo Faggioli. Frontiers of information access experimentation for research and education (Dagstuhl Seminar 23031). 2023. URL https://arxiv.org/abs/2305.01509.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623. Association for Computing Machinery, 2021.

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. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50. Association for Computing Machinery, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang

[11] https://waseda.box.com/ntcir17panel

[12] GFR quantifies the *expected user experience* for web search users, where the experience for each user group is defined based on the utility of the search engine result page as well as the similarity between the achieved and gold distributions over groups from a given attribute set (or more). It generalises the normalised cumulative utility (NCU) of Sakai and Robertson [2008], which quantifies the *expected user utility* over a population of users who issued the same query.

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. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment, 2023. URL https://arxiv.org/abs/2308.05374.

Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas4, and Svitlana Vakulenko. 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)*. NIST, 2023. URL https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf.

Tetsuya Sakai. Fairness-based evaluation of conversational search: A pilot study. In *Proceedings of EVIA 2023*, 2023a.

Tetsuya Sakai. On a few responsibilities of (IR) researchers (fairness, awareness, and sustainability): A keynote at ECIR 2023. *SIGIR Forum*, 57(1), 2023b. URL https://sigir.org/wp-content/uploads/2023/07/p04.pdf.

Tetsuya Sakai. SWAN: A generic framework for auditing textual conversational systems, 2023c. URL https://arxiv.org/abs/2305.08290.

Tetsuya Sakai and Stephen E. Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA 2008*, pages 30–41, 2008. URL http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/07-EVIA2008-SakaiT.pdf.

Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Trans. Inf. Syst.*, 42(1), 2023. URL https://dl.acm.org/doi/pdf/10.1145/3589763.

Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, reuse, recycle: Green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2825–2837. Association for Computing Machinery, 2022.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373, 2000.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, 2023. URL https://arxiv.org/abs/2309.10621.

Guido Zuccon, Harrisen Scells, and Shengyao Zhuang. Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 283–289. Association for Computing Machinery, 2023.