Overview of Papers Contributed to BREV-RAG@SIGIR-AP 2025

Tetsuya Sakai[1,2], Sijie Tao[1], Zhicheng Dou[3], Junjie Wang[4], Haoxiang Shi[5], Nuo Chen[6] and Atsuhi Keyaki[7]

[1]*Waseda University, Japan*
[2]*Naver Corporation, Korea*
[3]*Renmin University of China, P.R.C.*
[4]*Tsinghua University, P.R.C.*
[5]*Inner Mongolia University of Technology, P.R.C.*
[6]*The Hong Kong Polytechnic University, Hong Kong*
[7]*Hitotsubashi University, Japan*

Abstract

This paper provides a brief overview of five papers contributed to the BREV-RAG (Beyond Relevance-based EValuation of RAG systems) workshop, which will be held at SIGIR-AP 2025 on December 10, 2025. While RAG systems are usually evaluated in terms of relevance, correctness, and/or groundedness, the BREV-RAG workshop solicited papers that discuss possible evaluation axes other than the above that are important for making RAG more practical. Each submitted paper was reviewed by 2-4 program committee members (with at least one non-organiser PC member per paper), and we have accepted five papers. The papers will be presented in person in the first 90-minute session of the BREV-RAG workshop, and will be utilised as seed materials for group discussions in the second 90-minute session. The outcomes of the discussions will be reported in 2026 in SIGIR Forum (June 2026 issue).

Keywords

conversational search, evaluation, relevance, retrieval-augmented generation

1. Introduction

This paper provides a brief overview of five papers contributed to the BREV-RAG (Beyond Relevance-based EValuation of RAG systems) workshop, which will be held at SIGIR-AP 2025 on December 10, 2025 [1]. While RAG systems are usually evaluated in terms of relevance, correctness, and/or groundedness, the BREV-RAG workshop solicited papers that discuss possible evaluation axes other than the above that are important for making RAG more practical. As an example, the BREV-RAG call for papers referred to a set of evaluation axes presented in Sakai [2]; the table is duplicated in Table 1. Following his definitions, in the context of RAG, relevance (which he equates with coherence) means that the system response "looks like a plausible answer" given the user's question; correctness means that the system's response actually provides an factually accurate answer; and groundedness means that the system's claim is supported by a retrieved passage (or a combination of retrieved passages).[1]

For the BREV-RAG workshop, each submitted paper was reviewed by 2-4 program committee members (with at least one non-organiser PC member per paper)[2] in October 2025, and we have

BREV-RAG 2025: Beyond Relevance-based EValuation of RAG Systems, a SIGIR-AP 2025 workshop, held on December 10, 2025 in Xi'an, China.

tetsuyasakai@acm.org (T. Sakai); tsjmailbox@ruri.waseda.jp (S. Tao); dou@ruc.edu.cn (Z. Dou); wjj1020181822@toki.waseda.jp (J. Wang); hollis.shi@toki.waseda.jp (H. Shi); pleviumtan@toki.waseda.jp (N. Chen); a.keyaki@r.hit-u.ac.jp (A. Keyaki)
https://sakailab.com/tetsuya/ (T. Sakai)
0000-0002-6720-963X (T. Sakai); 0000-0002-6751-5303 (S. Tao); 0000-0002-9781-948X (Z. Dou); 0000-0001-9869-7085 (J. Wang); 0009-0002-9204-0351 (H. Shi); 0000-0001-8600-8203 (N. Chen); 0000-0001-6495-117X (A. Keyaki)

[1]While it may not be necessary to distinguish between relevance and correctness in the context of RAG response evaluation, according to Sakai's definition, abstention responses such as "*I cannot answer that question*" are relevant but not correct.
[2]One paper received only two reviews.

Table 1
20(+1) evaluation axes duplicated from Sakai [2].

	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?

accepted five papers. The papers will be presented in person in the first 90-minute session of the BREV-RAG workshop, and will be utilised as seed materials for group discussions in the second 90-minute session. The outcomes of the discussions will be reported in 2026 in SIGIR Forum.

The remainder of this paper will provide a brief overview of the five accepted papers, with comments on how they are related to each other.

2. Retrieval-Augmented Relevance Judgment for Specialized Domains

Authors: Taichi Motegi, Makoto P. Kato, Kazuhisa Hatakeyama, and Yosuke Yurikusa

The authors interpret the theme of our workshop "beyond relevance" as "beyond relevance assessment without domain-specific knowledge" and propose an approach to LLM-based relevance assessment that makes the LLM "more informed" about the search topic in advance; Their approach, called Retrieval-Augmented Relevance Judgement (RAJ), provides LLM with some relevant expert knowledge retrieved from a domain-specific corpus prior to the assessment step so that more reliable relevance labels can be obtained. Through experiments with data such as TREC-COVID, the authors show that their approach shows promise for specialised domains. In principle, their approach should work in any specialised domain.

3. RAG System for Supporting Japanese Litigation Procedures: Faithful Response Generation Complying with Legal Norm

Authors: Yuya Ishihara, Atsushi Keyaki, Hiroaki Yamada, Ryutaro Ohara, and Mihoko Sumida

The authors discuss several practical challenges when applying RAG to Japanese medical litigation procedures. In terms of "beyond relevance" evaluation axes, the authors argue that (a) the knowledge

sources utilised by the RAG system must be *compliant* in the context of Japanese litigation procedures;, (b) *faithfulness* (or *groundedness* from Table 1) is of utmost importance; and that (c) sources from *appropriate* timestamps must be utilised (where the appropriateness depends on "types of issues raised in trial"). We note that the third point is different from the definition of *adaptability* in Table 1, which is about keeping the information fresh; in the application discussed by Ishihara et al., identifying the relevant time period (rather than locating the most up-to-date information) is important.

4. Aspect-based Evaluation of Personalization and Diversification in Conversational Search Systems

Authors: Risa Hiramatsu, Rikiya Takehi, and Tetsuya Sakai

The authors propose an evaluation framework for quantifying how *personalised* (with respect to a given user profile) or *diversified* system responses in a conversation are. They define aspects as various facets or answers for a given topic within a conversation, and construct measures that utilise the aspects as building blocks. In their pilot experiment that reuses data from the TREC iKAT 2024 track [3], the authors utilise ChatGPT o4-mini to identify topics and aspects within the conversations in order to compute their personalisation and diversification measure scores. They observe a trade-off relationship between these two measures.

Customisability/Personalisability in Table 1 seems relevant to their work.

5. Measuring Group Fairness Differences in RAG Pipelines: A Pilot Study

Authors: Sijie Tao and Tetsuya Sakai

This study builds on the work of Tao et al. [4] and Sakai et al. [5], who evaluated the *group fairness* of conversational search responses. Unlike these existing studies, the authors quantify the *change* in group fairness across the retrieval (R) and generation (G) steps of RAG. By reusing runs from the NTCIR-17 FairWeb-1 web search subtask [6] and obtaining "pseudo-RAG" results, the authors demonstrate using their proposed measures that a RAG system's group fairness may increase or decrease as it moves from the R stage to the G stage.

Fair exposure in Table 1 is relevant to their work. The approach of Tao et al. makes one wonder: can we similarly quantify changes across the R and G steps in terms of other evaluation axes? For example, a RAG system may retrieve a lot of potentially *harmful* information at the R stage, but then successfully manage to provide a completely *harmless* response to the user after the G stage.

6. Evaluating the Modesty of RAG Systems at the NTCIR-19 R2C2 Task: Potential Challenges

Tetsuya Sakai, Sijie Tao, Atsuya Ishikawa, Hanpei Fang, Ziliang Zhao, Yujia Zhou, Nuo Chen, and Young-In Song

The authors describes the ongoing NTCIR-19 R2C2 (RAG Responses: Confident and Correct?) task and its challenges.[3] The task evaluates RAG responses not only in terms of correctness but also in terms of *modesty* (See Table 1), i.e., whether the system's confidence about its own response aligns with correctness. To distinguish between *overconfidence* (i.e., high confidence for an incorrect response) and *underconfidence* (i.e., low confidence for a correct response) and to encourage systems to strike a balance between the two, the task employs the HMR (Harmonic Mean of Rewards) evaluation framework [7]. If confidence scores of RAG systems are appropriate, then downstream tasks will be able to decide

[3] http://sakailab.com/r2c2/

whether they should utilise the answer from the RAG system or not. Moreover, if the confidence scores are appropriately low for unreliable answers, systems can say "*I don't know.*" when it should.

7. Summary and Future Work

This paper provided a brief overview of the five papers that have been accepted for the BREV-RAG workshop at SIGIR-AP 2025. Please check out the June 2026 issue of SIGIR Forum to see the summary of our group discussions held at the workshop.

The current edition of BREV-RAG covered only a small number of axes that "go beyond relevance, correctness, and groundedness." Should there be need from the SIGIR(-AP) community, we hope to organise a second BREV-RAG workshop in the near future.

Declaration on Generative AI

The authors did not use any generative AI tools for writing this paper.

Acknowledgement

We thank the following non-organiser PC members for their constructive feedback on the submitted papers: Maria Maistro (University of Copenhagen), Marwah Alaofi (RMIT University), Yuto Nakachi (University of Tsukuba), Negar Arabzadeh (University of California), Fabrizio Silvestri (University of Rome), Giovanni Trappolini (University of Rome), Federico Siciliano (University of Rome), Mohammad Aliannejadi (University of Amsterdam), Nandan Thakur (University of Waterloo).

References

[1] T. Sakai, S. Tao, Z. Dou, J. Wang, H. Shi, N. Chen, A. Keyaki, BREV-RAG: Beyond relevance-based EValuation of RAG systems: A SIGIR-AP 2025 workshop proposal, in: Proceedings of ACM SIGIR-AP 2025, 2025, p. to appear.

[2] T. Sakai, SWAN: A generic framework for auditing textual conversational systems, 2023. arXiv:2305.08290.

[3] M. Aliannejadi, Z. Abbasiantaeb, S. Lupart, S. Chatterjee, J. Dalton, L. Azzopardi, TREC iKAT 2024: The interactive knowledge assistance track overview, in: Proceedings of TREC 2024 (NIST SP1329), 2025. URL: https://trec.nist.gov/pubs/trec33/papers/Overview_ikat.pdf.

[4] S. Tao, T. Sakai, J. Wang, H. Fang, Y. Zhang, H. Li, Y. Tu, N. Chen, M. Maistro, Overview of the NTCIR-18 FairWeb-2 task, in: Proceedings of NTCIR-18, 2025, pp. 40–60. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings18/pdf/ntcir/01-NTCIR18-OV-FAIRWEB-TaoS.pdf.

[5] T. Sakai, S. Tao, Y.-I. Song, Evaluating group fairness and relevance in conversational search: An alternative formulation, in: Proceedings of EVIA 2025, 2025, pp. 15–22. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings18/pdf/evia/03-EVIA2025-EVIA-SakaiT.pdf.

[6] S. Tao, N. Chen, T. Sakai, Z. Chu, H. Arai, I. Soboroff, N. Ferro, M. Maistro, Overview of the NTCIR-17 FairWeb-1 task, in: Proceedings of NTCIR-17, 2023, pp. 284–305. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings17/pdf/ntcir/01-NTCIR17-OV-FAIRWEB-TaoS.pdf.

[7] T. Sakai, Evaluating system responses based on overconfidence and underconfidence, in: Joint Proceedings of the SIGIR-AP 2024 Workshops EMTCIR 2024 and UM-CIR 2024, 2024. URL: https://ceur-ws.org/Vol-3854/emtcir-1.pdf.