Report on the 1st Workshop on Beyond Relevance-Based EValuation of RAG Systems (BREV-RAG 2025) at SIGIR-AP 2025

Tetsuya Sakai
Waseda University/Naver Corporation
Japan
tetsuyasakai@acm.org

Hanpei Fang
Waseda University
Japan
hanpeifang@ruri.waseda.jp

Makoto P. Kato
University of Tsukuba
Japan
mpkato@acm.org

Keping Bi, Hideo Joho, Charles L.A. Clarke, Sijie Tao, Junjie Wang, Haoxiang Shi, Risa Hiramatsu, Hengran Zhang, Jiaxin Mao, Yuya Ishihara, Taichi Motegi, Nuo Chen, Zhicheng Dou, Atsushi Keyaki *

Abstract

We report on the outcomes of the first BREV-RAG (Beyond Relevance-based EValuation of RAG systems) workshop, a half-day post-conference workshop of ACM SIGIR-AP 2025. The workshop featured five paper presentations and breakout group discussions. After the paper presentations, three evaluation axes/themes were chosen by the workshop participants and accordingly three discussion groups were formed, namely, *modesty*, *diversity and fairness*, and *specialised domains* groups. We report on what each of the groups discussed during and after the workshop.

Date: 10 December 2025.

Website: http://sakailab.com/brev-rag/.

1 Introduction

RAG (Retrieval-Augmented Generation) combines the parametric general knowledge of large language models with information retrieved from domain-specific, easily updatable corpora, and enables natural language-based conversation with users. While it has quickly become a commodity, many open questions remain as to how we should evaluate RAG responses, especially from viewpoints other than whether the responses is correct or not or whether the response is faithful to the retrieved information. Thus, the purpose of the BREV-RAG workshop was to provide an opportunity for SIGIR-AP 2025 participants to get together for short but focussed discussions on evaluating RAG systems using axes other than just relevance, correctness, or *groundedness* (i.e., "is the response generated by the G part of the RAG backed up by information retrieved by the R part?").

*Affiliation not shown for all authors due to space limitations (see Appendix B for details).

The first 90 minutes of the workshop consisted of presentations of five refereed and accepted papers. We thank the program committee members; they are listed in Appendix A. The BREV-RAG paper proceedings can be found online.[1]

The second 90 minutes of the workshop was for breakout discussions. Three evaluation axes/themes were chosen by the workshop participants through voting: *modesty* (based on a presentation by Sakai et al. [2025]), *diversity and fairness* (based on presentations by Hiramatsu et al. [2025] and Tao and Sakai [2025]), and *specialised domains* (based on presentations by Ishihara et al. [2025] and Motegi et al. [2025]). Accordingly, three discussion groups were formed. We report on what each of the groups discussed during and after the workshop.

2 Modesty (Overconfidence and Underconfidence)

Contributors: Keping Bi, Hideo Joho, Charles L.A. Clarke, and Tetsuya Sakai (group leader)

Calibration is the task of aligning the system's confidence with its actual accuracy. However, in traditional calibration, the difference between *overconfidence* (i.e., the system is confident about its wrong answer) and *underconfidence* (i.e., the system is unconfident about its correct answer) is not taken into account. While LLMs tend to be overconfident about its own responses, underconfidence can also be a potential problem, as users or downstream tasks will not be able to rely on the correct answers if the system lacks confidence. Thus, if tuning a system to supress its tendency to be overconfident results in more underconfidence, this would not be what we want. By *modesty* evaluation, we mean trying to suppress both overconfidence and underconfidence [Sakai, 2024]. The ongoing NTCIR-19 R2C2 (RAG Responses: Confident and Correct?) task features modesty evaluation [Sakai et al., 2025]. Note that, if a system's confidence score about its candidate answer is *appropriately low*, it can choose to respond with "*I don't know*," "*I don't have enough confidence to answer that question*," etc.

In the Modesty Group, the following were discussed.

- The R2C2 task is a highly collaborative task, in that the AC (Answering with Confidence) subtask participants (responsible for the G part of RAG) can utilise any of the PR (Passage Retrieval) subtask runs (responsible for the R part of RAG). Here, a set of relevant passages $\{P_1, P_2\}$ may be useful for AC system 1 but not so for AC system 2, while another set of relevant passages $\{P_3, P_3\}$ may be useful for AC system 2 but not so for AC system 1. Synergy across the two subtasks should be studied.
- How should participants come up with answer confidence scores? If a set of returned relevant nuggets $\{A, B\}$ can derive the correct answer *and* answer set of returned relevant nuggets $\{A, C, D\}$ can also do the same,[2] should the answer be given a higher confidence score than when it is backed up by only one set of nuggets? How should the participants utilise the passage ranks from the PR subtask, or the "reputation" of the PR subtask participants? (Current answer from an R2C2 task organiser: these questions are left for the participants to explore!)

[1] https://brev-rag-workshop.github.io/

[2] Sakai [2019, Footnote 13] details the idea of evaluation by *combinational relevance*.

- A possible variant of the current R2C2 task: in the input question, insert a remark such as "*There might not be a correct answer in the corpus*" and study the effect on the returned confidence scores.
- Another possible variant of the task is to make systems return confidence expressions in natural language. (Comment from an R2C2 task organiser: even in that case, the system should probably have an appropriate internal confidence score as a number, so that it can choose to respond with "*I don't know*" etc. when it should do so.)
- A general question: when should the G part of the RAG trust its parametric knowledge over the information returned by the R part when the two contradict with each other? One way to measure the *contextual faithfulness* [Ming et al., 2025] of RAG would be to inject counterfactual passages at the R stage and see how the G part handles them. If the G part completely trusts the counterfactual passages when generating its answer, then the G part is faithful. (For the R2C2 task, the AC subtask participants are expected to "trust" the retrieved passages: there will be no deliberately injected counterfactual passages. However, the AC runs may potentially generate counterfactual *nuggets* from the (supposedly factual) retrieved passages, which are called *bogus* nuggets in the R2C2 task.)
- There are some concepts that are related to modesty, overconfidence, and underconfidence. In Askell et al. [2021], "calibratedness" is listed as one of the aspects of *honesty*; however, recall that this ignores the distinction between overconfidence and underconfidence. Moreover, *sycophancy* [Liu et al., 2024b] is a know problem with LLMs: they try to say what the user wants just to please the user. Hence, LLM-based system might change their confidence scores in response to user remarks such as "*Are you absolutely sure?*" Change in system confidence across multiple conversational turns is outside the scope of the ongoing R2C2 task.

3 Diversity and Fairness

Contributors: Sijie Tao, Junije Wang, Haoxiang Shi, Risa Hiramatsu, and Hanpei Fang (group leader)

Diversity (accommodating different search engine user needs with a single SERP) and *fairness* (providing fair exposure to different entities that are potential search targets) have long been important evaluation axes in Information Retrieval. With the rise of RAG systems, their importance has been amplified. In traditional IR systems, these issues primarily affect what users can *see* on a search engine results page (SERP), where they can inspect multiple documents. RAG systems, however, usually provide a single, information-dense explanatory answer. Prior work suggests that users tend to trust such answers [Steyvers et al., 2025], which implies that these two issues have greater potential to directly shape what users *believe*, especially for complex information needs. As a result, insufficient diversity or unfairness in retrieval and generation can systematically narrow viewpoints or disproportionately disadvantage certain groups at scale, making the evaluation of diversity and fairness essential for building trustworthy RAG systems.

In the Diversity and Fairness group, the following topics were discussed:

- Why *diversity* and *fairness* matter in RAG: Unlike SERP-based systems, where users may inspect multiple documents, RAG systems often provide a single, information-dense ex-

planatory answer. Since users tend to trust such answers [Steyvers et al., 2025], diversity and fairness issues in both retrieval and generation may more directly shape what they *believe* [White, 2013].

- Diversity is closely related to *query ambiguity* and *search intents*. In traditional web search, many queries are short and underspecified, and even question-form queries can remain ambiguous. For example, "What is SIGIR?" typically refers to the ACM Special Interest Group on Information Retrieval, but it can also refer to other entities such as a logistics company in Russia. Since RAG systems generate synthesized overview answers, intent coverage and viewpoint balance become particularly important for such ambiguous queries.
- In traditional IR, fairness concerns are often discussed in terms of SERP exposure, where users can inspect multiple documents and cross-check information across different sources. Since each document typically contributes only a small slice of evidence, users can compare results, identify inconsistencies, and gradually build their own understanding. In contrast, RAG often presents a single information-dense "overview" answer. This presentation reduces opportunities and incentives to verify claims against alternative sources, question the framing, or notice missing evidence; consequently, even small retrieval or generation biases may be amplified. As a result, unfair or biased retrieval/generation may more directly shape what users *believe*, not merely what they *see*.
- Human-centric SERP evaluation metrics may not transfer cleanly to the R part of RAG systems [Tao and Sakai, 2025]. A ranked list that is "good" for a human reader may not be "good" for an LLM generator. Many human-centric metrics inherently assume that users stop early once they are satisfied and therefore emphasize top-ranked documents. In contrast, an LLM generator generally consumes *all* the input tokens it is given, and it can be sensitive to ordering and surface-level presentation cues, exhibiting positional bias (e.g., "Lost in the Middle" [Liu et al., 2024a]) as well as cognitive biases (e.g., threshold priming [Chen et al., 2024] and recency bias [Fang et al., 2025]). Therefore, using such metrics without adjustment may misestimate the utility of a ranked list for the G component of RAG, highlighting the need for generator-aware metrics to evaluate SERPs produced by the R component.
- A practical issue is how to implement the evaluation. The group discussed utilizing existing test collections, including data from the NTCIR INTENT tasks [Song et al., 2011; Sakai et al., 2013] for retrieval diversity and from the NTCIR FairWeb tasks [Tao et al., 2023, 2025] for retrieval fairness, or extending them to better reflect RAG characteristics. Then, it is necessary to develop retrieval-side metrics whose scores correlate with (and ideally predict) the diversity/fairness observed in the generated responses, so that SERPs can be rewarded based on their expected downstream behaviour.
- While the inputs to the generator are controllable, the randomness of LLM outputs can be an issue, which means that a single-run evaluation might be unreliable. A practical approach discussed was to generate multiple responses per query and to evaluate the agreement and/or the distribution across them.

4 Specialised Domains

Contributors: Hengran Zhang, Jiaxin Mao, Yuya Ishihara, Taichi Motegi, and Makoto P. Kato (group leader)

In this group, we discussed evaluation metrics that become particularly important when RAG is deployed in specialised domains such as legal, finance, and medical. Unlike general-purpose settings, the key concern is not only whether the answer is correct, but whether the system can support safe decisions under strict reliability requirements, where even a small number of severe failures may be unacceptable.

In the Specialised Domains Group, the following were discussed.

- *Risk of decision.* In specialised domains, RAG outputs can be used to guide or justify actions (e.g., compliance decisions, financial judgments, and clinical advice). A single harmful or misleading suggestion can lead to irrevocable damage. Thus, we discussed *risk of decision* as an evaluation metric: the system should ideally emit a risk score (or risk level) alongside the answer, so that users can appropriately calibrate their reliance on the output, especially in borderline cases. This metric is closely related to modesty, but they capture different aspects of reliability:
  - **Modesty**: how confident the system is that its own answer is correct.
  - **Risk of decision**: how severe the practical impact would be if its own answer is incorrect.

- *Faithfulness.* Faithfulness was viewed as non-negotiable in critical applications because users (or experts in specialised domains) must be able to verify the answer against retrieved evidence. If the answer diverges from the retrieved documents, the system does not reduce user burden; instead, it shifts effort to manual verification and increases the chance of erroneous decisions. Hence, faithfulness is a core requirement for making RAG outputs operationally usable especially in specialised domains, though this metric has been widely used and discussed in general RAG systems.
- *Misinformation ratio.* While misinformation is conceptually related to answer correctness, we treated it as a distinct factor that warrants heavier penalties in specialised domains. In ordinary domains, small factual errors may be tolerated; in specialised domains, non-factual statements can cause critical problems. Therefore, misinformation-sensitive evaluation should explicitly penalise non-factual content, rather than averaging it away under general accuracy metrics.
- *Robustness.* We discussed robustness not as the average level of quality, but as the frequency of critical failures: i.e., how often the system produces risky, unfaithful, or wrong outputs across a query distribution. This framing matters because in high-stakes environments, a few unacceptable cases cannot be allowed, even if average performance appears strong. This notion of robustness, or *minimizing the probability of significant failure*, was central to the *risk-sensitive retrieval task* in the TREC 2013 Web Track [Collins-Thompson et al., 2013]. Risk-sensitive evaluation measures could be adapted to RAG evaluation in specialised domains.
- *Recall.* We noted that recall can be an important metric when the user's goal is comprehensive coverage, such as generating a domain survey, identifying all relevant legal events,

collecting financial risk factors, or enumerating possible symptoms. Under such tasks, low coverage can itself be harmful: missing a key precedent, a critical financial signal, or a clinically relevant symptom can lead to wrong downstream judgments.

- *Time-sensitiveness.* Specialised domains often contain rapidly changing knowledge: laws can be amended, medical guidance can be updated, and financial facts can become obsolete. Therefore, we discussed time-sensitiveness as an evaluation metric: a system should be judged on whether it provides valid information at the time of the query, not merely information that was once correct. This suggests that evaluation datasets and protocols should account for temporal validity and document freshness.
- *Difficulty level.* We also discussed that specialised-domain corpora are frequently written for experts, while RAG users may range from experts to non-experts. Hence, beyond factual and evidence-related metrics, the difficulty level of the answer can be important: experts may prefer concise, technical statements, while non-experts may require accessible explanations aligned with their knowledge level.

Finally, we hypothesised that which metrics matter most depends strongly on the corpus. Two corpus-level characteristics were highlighted in our discussion:

- *Misinformation ratio* (e.g., social media-like corpora vs. curated textbooks/laws), which changes the necessity of misinformation- and risk-focused evaluation.
- *Dynamics of validity* (i.e., how quickly documents become outdated), which determines how critical time-sensitiveness should be in evaluation.

5 Summary

We reported on what happened at the first BREV-RAG workshop of the SIGIR-AP 2025 conference, held in December 2025. More specifically, each of the three breakout groups (on *modesty*, *diversity/fairness*, and *specialised domains*) provided a summary of their discussions. The NTCIR-19 R2C2 task explores modesty evaluation for RAG systems; it is hoped that RAG will also be evaluated along other important evaluation axes through well-designed shared tasks in the near future. Other existing "beyond relevance/correctness/groundedness" RAG evaluation include: Detection, Retrieval, and Augmented Generation for Understanding News (DRAGUN) at TREC (*trustworthiness*)[3] and Advertisement in Retrieval-Augmented Generation of Touché at CLEF@2026 (advertisement detection and removal).[4]

A Program Committee

- Tetsuya Sakai, Waseda University (workshop organiser)
- Sijie Tao, Waseda University (workshop organiser)
- Zhicheng Dou, Renmin University of China (workshop organiser)
- Junjie Wang, Tsinghua University (workshop organiser)

[3] https://trec-dragun.github.io/

[4] https://touche.webis.de/clef25/touche25-web/advertisement-detection.html

- Haoxiang Shi, Inner Mongolia University of Technology (workshop organiser)
- Nuo Chen, The Hong Kong Polytechnic University (workshop organiser)
- Atsushi Keyaki, Hitotsubashi University (workshop organiser)
- 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
- Mohammad Aliannejadi, University of Amsterdam
- Nandan Thakur, University of Waterloo

B Authors and Affiliations

First author tier:

- Tetsuya Sakai, Waseda University/Naver Corporation, Japan.
- Hanpei Fang, Waseda University, Japan.
- Makoto P. Kato, University of Tsukuba, Japan.

Second author tier:

- Keping Bi, Chinese Academy of Sciences, China.
- Hideo Joho, University of Tsukuba, Japan.
- Charles L.A. Clarke, University of Waterloo, Canada.
- Sijie Tao, Waseda University, Japan.
- Junjie Wang, Tsinghua University, China.
- Haoxiang Shi, Inner Mongolia University of Technology, China.
- Risa Hiramatsu, Waseda University, Japan.
- Hengran Zhang, Chinese Academy of Sciences, China.
- Jiaxin Mao, Renmin University of China, China.
- Yuya Ishihara, Hitotsubashi University, Japan.
- Taichi Motegi, University of Tsukuba, Japan.
- Nuo Chen, The Hong Kong Polytechnic University, China.
- Zhicheng Dou, Renmin University of China, China.
- Atsushi Keyaki, Hitotsubashi University, Japan.

References

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. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. Ai can be cognitively biased: An exploratory study on threshold priming in llm-based batch relevance assessment. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 54–63, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707247. doi: 10.1145/3673791.3698420. URL https://doi.org/10.1145/3673791.3698420.

Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. TREC 2013 web track overview. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013*, 2013.

Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. Do large language models favor recent content? a study on recency bias in llm-based reranking. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, page 85–94, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722189. doi: 10.1145/3767695.3769493. URL https://doi.org/10.1145/3767695.3769493.

Risa Hiramatsu, Rikiya Takehi, and Tetsuya Sakai. Aspect-based evaluation of personalization and diversification in conversational search systems. In *Proceedings of the International Workshop on Beyond Relevance-based EValuation of RAG Systems (BREV-RAG) 2025*, pages 22–33, 2025. URL https://brev-rag-workshop.github.io/paper-03.pdf.

Yuya Ishihara, Atsushi Keyaki, Hiroaki Yamada, Ryutaro Ohara, and Mihoko Sumida. RAG system for supporting japanese litigation procedures: Faithful response generation complying with legal norms. In *Proceedings of the International Workshop on Beyond Relevance-based EValuation of RAG Systems (BREV-RAG) 2025*, pages 59–65, 2025. URL https://brev-rag-workshop.github.io/paper-06.pdf.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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, 2024b. URL https://arxiv.org/abs/2308.05374.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows", 2025. URL https://arxiv.org/abs/2410.03727.

Taichi Motegi, Makoto P. Kato, Kazuhisa Hatakeyama, and Yosuke Yurikusa. Retrieval-augmented relevance judgment for specialized domains. In *Proceedings of the International Workshop on Beyond Relevance-based EValuation of RAG Systems (BREV-RAG) 2025*, pages 34–46, 2025. URL https://brev-rag-workshop.github.io/paper-04.pdf.

Tetsuya Sakai. Graded relevance assessments and graded relevance measures of NTCIR: A survey of the first twenty years. *CoRR*, abs/1903.11272, 2019. URL http://arxiv.org/abs/1903.11272.

Tetsuya 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.

Tetsuya Sakai, Zhicheng Dou, Takehiro Yamamoto, Yiqun Liu, Min Zhang, Ruihua Song, MP Kato, and M Iwata. Overview of the ntcir-10 intent-2 task. In *NTCIR*, 2013.

Tetsuya Sakai, Sijie Tao, Atsuya Ishikawa, Hanpei Fang, Ziliang Zhao, Yujia Zhou, Nuo Chen, and Young-In Song. Evaluating the modesty of RAG systems at the NTCIR-19 R2C2 task: Potential challenges. In *Proceedings of the International Workshop on Beyond Relevance-based EValuation of RAG Systems (BREV-RAG) 2025*, pages 5–21, 2025. URL https://brev-rag-workshop.github.io/paper-02.pdf.

Ruihua Song, Min Zhang, Tetsuya Sakai, Makoto P Kato, Yiqun Liu, Miho Sugimoto, Qinglei Wang, and Naoki Orii. Overview of the ntcir-9 intent task. In *NTCIR*, 2011.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Sijie Tao and Tetsuya Sakai. Measuring group fairness differences in RAG pipelines: A pilot study. In *Proceedings of the International Workshop on Beyond Relevance-based EValuation of RAG Systems (BREV-RAG) 2025*, pages 47–58, 2025. URL https://brev-rag-workshop.github.io/paper-05.pdf.

Sijie Tao, Nuo Chen, Tetsuya Sakai, Zhumin Chu, Hiromi Arai, Ian Soboroff, Nicola Ferro, and Maria Maistro. Overview of the ntcir-17 fairweb-1 task. 2023.

Sijie Tao, Tetsuya Sakai, Junjie Wang, Hanpei Fang, Yuxiang Zhang, Haitao Li, Yiteng Tu, Nuo Chen, and Maria Maistro. Overview of the ntcir-18 fairweb-2 task. 2025.

Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 3–12, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320344. doi: 10.1145/2484028.2484053. URL https://doi.org/10.1145/2484028.2484053.