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

Tetsuya Sakai[1,2], Sijie Tao[1], Atsuya Ishikawa[1], Hanpei Fang[1], Ziliang Zhao[3], Yujia Zhou[4], Nuo Chen[5] and Young-In Song[2]

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

Abstract

If a conversational search system or a RAG (Retrieval-Augmented Generation) system can return an appropriate confidence score along with its answer, this would be informative for the user and the downstream tasks. Moreover, if its own confidence score is low, the system itself may opt to say "I don't know" instead of returning the answer. Ideally, the confidence score should align with the answer accuracy: it should be high for correct answers and low for incorrect answers. Sakai refers to this property as *modesty.*

The ongoing NTCIR-19 R2C2 (RAG Responses: Confident and Correct?) task, which is to be concluded in December 2026, implements the evaluation of RAG systems' modesty. This workshop describes the task design of R2C2, and discusses possible challenges and future work.

Keywords

calibration, confidence, conversational search, evaluation, retrieval-augmented generation

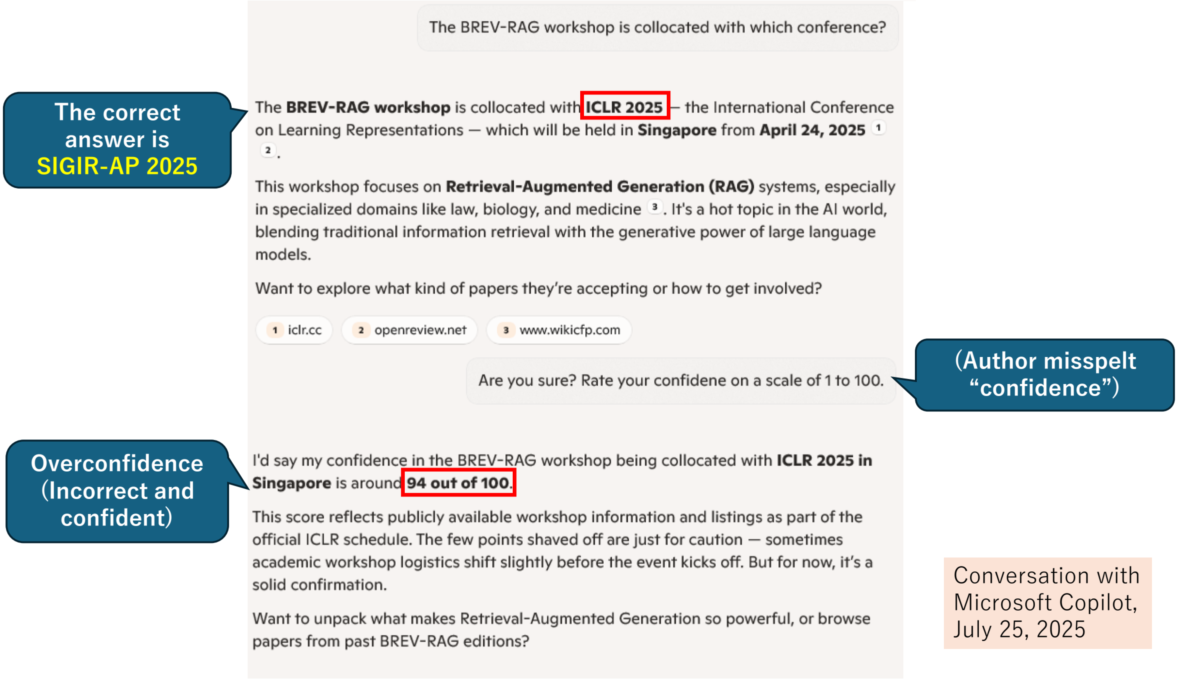
1. Introduction

Large language models (LLMs) *hallucinate* [1, 2], often with confidence (e.g., [3, Figure 9]). Figure 1 is our own example: even though the system returned the wrong answer ("ICLR 2025"), it says that it is 94% confident about it. We say that a system is *overconfident* if its confidence score is large for an incorrect answer, and that it is *underconfident* if its confidence score is small for a correct answer. Note that underconfidence is also a potential problem: if the system is unsure about its answer, the user or the downstream task cannot rely on or utilise the answer. Clearly, the confidence score should align with the answer's actual accuracy: if the confidence score is appropriately low for an incorrect answer, this will let the system opt to return "*I don't know*" (IDK) instead of that answer [2, 4].

While the task of aligning confidence scores with accuracy is known as *calibration* [5], traditional calibration measures do not distinguish between overconfidence and underconfidence. Hence, for example, they cannot let researchers quantify how often an LLM suffers from overestimation rather than underestimation. In light of this, Sakai [6] proposed a simple evaluation measure called HMR (Harmonic Mean of Rewards), which, unlike existing calibration measures, evaluates overconfidence and underconfidence separately and is guaranteed to (a) penalise a lowered confidence for a correct response; *and* (b) penalise a raised confidence for an incorrect response. To distinguish the evaluation based on this approach from traditional calibration evaluation, we shall refer to the former as *modesty*

This paper was reviewed and accepted by the program committee for 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); shiro0406@ruri.waseda.jp (A. Ishikawa); hanpeifang@ruri.waseda.jp (H. Fang); zhaoziliang@ruc.edu.cn (Z. Zhao); zhouyujia@mail.tsinghua.edu.cn (Y. Zhou); pleviumtan@toki.waseda.jp (N. Chen); song.youngin@navercorp.com (Y. Song)
https://sakailab.com/tetsuya/ (T. Sakai)
0000-0002-6720-963X (T. Sakai); 0000-0002-6751-5303 (S. Tao); 0009-0005-2823-5172 (A. Ishikawa); 0009-0009-5117-5621 (H. Fang); 0000-0001-8169-0143 (Z. Zhao); 0000-0002-3530-3787 (Y. Zhou); 0000-0001-8600-8203 (N. Chen); 0000-0003-0669-005X (Y. Song)



The BREV-RAG workshop is collocated with which conference?

The **BREV-RAG workshop** is collocated with **ICLR 2025** — the International Conference on Learning Representations — which will be held in **Singapore** from **April 24, 2025** [1] [2].

This workshop focuses on **Retrieval-Augmented Generation (RAG)** systems, especially in specialized domains like law, biology, and medicine [3]. It's a hot topic in the AI world, blending traditional information retrieval with the generative power of large language models.

Want to explore what kind of papers they're accepting or how to get involved?

1 iclr.cc 2 openreview.net 3 www.wikicfp.com

Are you sure? Rate your confidene on a scale of 1 to 100.

I'd say my confidence in the BREV-RAG workshop being collocated with **ICLR 2025 in Singapore** is around **94 out of 100**.

This score reflects publicly available workshop information and listings as part of the official ICLR schedule. The few points shaved off are just for caution — sometimes academic workshop logistics shift slightly before the event kicks off. But for now, it's a solid confirmation.

Want to unpack what makes Retrieval-Augmented Generation so powerful, or browse papers from past BREV-RAG editions?

The correct answer is SIGIR-AP 2025

(Author misspelt "confidence")

Overconfidence (Incorrect and confident)

Conversation with Microsoft Copilot, July 25, 2025

Figure 1: An example of LLM's overconfidence in its own response.

evaluation [7]. While Kalai et al. [2] have recently proposed to evaluate system responses including IDK responses in order to avoid penalising systems that can return IDK, their approach does not take the *degree* of uncertainty into account.

RAG (Retrieval-Augmented Generation) systems [8, 9] combine the strengths of computationally expensive but "intelligent" LLMs and domain-specific, easily-updatable, and fast passage retrieval techniques. RAG is receiving a lot of attention recently, as exemplified by the fact that TREC 2025 features as many as five RAG or RAG-related tasks (RAG, BioGen, DRAGUN, iKAT, and RAGTIME).[1] RAG systems are easily to implement and hence quickly becoming a commodity, making practical and social impacts in our lives. As we believe that the modesty evaluation of RAG systems is important but missing in existing studies, we proposed the R2C2 (RAG Responses: Confident and Correct?) task for NTCIR-19, which is to be concluded in December 2026; the task proposal was accepted in May 2025.

This paper describes the design of the NTCIR-19 R2C2 task. The task covers two subtasks, PR (Passage Retrieval) and AC (Answering with Confidence), and offers the following key features.

- In the PR subtask, passage relevance is automatically derived from the results of the AC evaluation;
- In the AC subtask, the answer confidence score is evaluated along with the answer accuracy, by leveraging HMR;
- AC subtask participants are encouraged to utilise PR runs from other participants, so that best practices from each subtask can potentially be combined;
- In the AC subtask, the evaluation is conducted automatically by leveraging LLMs (with some manual post-editing if necessary).

We also respond to some of the reviewer comments as we go along. In addition, this paper discusses possible challenges and future work for the task.

[1] https://trec-rag.github.io/#explore-more-rag-tracks-at-trec-2025

2. Prior Art

2.1. Going beyond Relevance for Conversational Search and RAG

In the context of evaluating the textual responses of conversational search and RAG systems, few studies have addressed the evaluation axes other than answer relevance, correctness, and groundedness. By *groundedness* we mean whether the answer has a proper supporting piece of evidence [7, 10]. A similar concept is *faithfulness*, which Wallat et al. [11] define as whether the answer is actually directly derived from the citation that a RAG system attaches to the answer (rather than *post-rationalised* after relying on the system's parametric knowledge). Askell et al. [12] define an "AI" as *aligned* (with human values) if it is *helpful, honest*, and *harmless* (in other words, *safe* [13]). Wang et al. [14] discuss *helpfulness* and *level of detail* of LLM responses (To be discussed more in Section 2.3). More evaluation axes for conversational search are discussed in Sakai [7, Table 1], Murugadoss et al. [15, Figure 2], etc. Chen et al. [16] propose an LLM-based muti-axes evaluation framework for conversational recommendation systems.

In the context of ranked retrieval (which is relevant to the "R" part of RAG), *fair exposure* has been evaluated as exemplified by the TREC Fair Rankings tracks [17] and the NTCIR FairWeb tasks [18]. Kim and Diaz [19] address *individual fairness* evaluation specifically for the "R" part of RAG. Also, bias evaluation for rankings (e.g., for addressing gender bias [20]) is closely related to fair ranking evaluation. The NTCIR-18 FairWeb task featured a conversational search subtask, where the textual output of a conversational search system (not a ranked list of documents, but a plain text that contains a list of entities) was evaluated based on *group fairness* [21, 22, 18]. Also, for evaluating generative IR [23], Gienapp et al. [24] present five evaluation criteria: *coherence, coverage, consistency, correctness*, and *clarity*.

2.2. Using Nuggets for Evaluating Textual Responses

The NTCIR-19 R2C2 task defines a *nugget* as a factual claim (derived from a retrieved passage) represented as a single sentence, and uses nuggets for semiautomatic evaluation of answer accuracy in the Answering with Confidence subtask.[2] Hence, this section provides a brief overview of nugget-based and related approaches to evaluating textual responses.

Before the advent of LLMs, the TREC 2023 Question Answering (QA) track implemented an evaluation method based on manually defined *vital* and *okay* nuggets [25]. Nenkova et al. [26] proposed to evaluate textual summaries based on manually defined *semantic content units*, which can also be seen as a form of nuggets. Nugget-based methods were also used for evaluating complex QA at TREC [27] and NTCIR [28]. Sakai et al. [29] proposed to evaluate textual responses based on nuggets and their *positions* within the response; Sakai and Dou [30], Kato et al. [31], and Zeng et al. [32] extended that approach for more complex textual evaluation tasks.

Nugget-based approaches appear to be regaining momentum after the advent of LLMs [33, 34], as LLMs now enable (semi)automatic nugget creation, nugget identification within a given text, etc., with reasonable accuracy. For example, Pradeep et al. [35] describe an LLM-based automatic approach to RAG evaluation for the TREC 2024 RAG track by "refactoring" the aforementioned nugget-based evaluation approach of the TREC 2003 QA track. Abbasiantaeb et al. [36] also describe an LLM-based, nugget-based RAG evaluation pipeline in the context of the TREC iKAT 2024 track. The TREC 2024 NeuCLIR track also utilises a nugget-based evaluation method [37]. Samarinas et al. [38] propose to evaluate the coverage of "atomic factual claims" for a generated response. More recently, Thakur et al. [39] have proposed a framework for constructing nugget-based IR and RAG benchmarks for technical documents.

[2] While a nugget is often defined as an *atomic* claim, our definition does not require atomicity as our view is that it is practically difficult to determine whether a claim is atomic or can be further decomposed. For example, is "*Sakai was born in 2000 in Osaka.*" atomic? If so, what about "*Sakai was born in 2000.*"? If the latter is atomic, then is the former really atomic?

2.3. Using LLMs for Evaluating Textual Responses

The NTCIR-19 R2C2 task employs LLMs for semiautomatic evaluation of the answers returned by RAG systems. Hence, this section provides a brief overview of LLM-based (semi)automatic evaluation approaches for textual responses.

Wang et al. [14] discuss the problems with "evaluating LLM responses using LLMs" in terms of *helpfulness* and *level of detail* as well as relevance and accuracy. Hashemi et al. [40] propose an LLM-based evaluation method that builds on LLMs' probability distributions for *rubric* questions, which can be similar to our axes such as correctness, groundedness, and conciseness; they report that their combined results align well with human assessment. See Li et al. [41] and Peng et al. [42] for surveys (as of 2024) on LLM-based evaluation of textual responses. Murugadoss et al. [15] discuss meta-evaluation of LLM-as-a-judge in this context. The aforementioned work by Chen et al. [16] propose an LLM-based evaluation method for conversational recommendation systems.

In the context of ranked retrieval evaluation, there is a controversy as to whether LLM-based relevance assessments can/should replace manual relevance assessments [43, 44, 45, 46]. Some of these arguments are also relevant to evaluating textual responses. For example, there is the question of *circularity*: if an LLM is intelligent enough to judge whether another LLM's answers is correct or not, why didn't we just use the former LLM as our answerer in the first place? Moreover, LLMs can easily be fooled [47], and can be *narcissistic* [43, 48]: that is, LLM-based evaluators tend to favour LLM-based systems.

Our standpoint regarding the above controversy is as follows: "*Yes LLMs are still too dumb to replace reliable humans in complex assessment and annotation tasks. However, they are quite trustworthy if we give them clearly-defined, simple tasks such as textual entailment assessment for a given pair short texts.*" The NTCIR-19 R2C2 task has been designed accordingly, as we shall discuss in Section 3.

2.4. Calibration Measures and HMR

Sakai [6] compared his HMR measure with existing calibration measures (*Expected Calibration Error* (ECE), *Maximum Calibration Error* (MCE) [5, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60], the Brier score [51, 58, 61, 62, 63], KS [64]) based on axiomatic considerations. He showed that while HMR satisfies his three axioms based on points (a) and (b) mentioned in Section 1 (that is, a lowered confidence for a correct response and a raised confidence for an incorrect response should both be penalised), none of the other measures are guaranteed to do so.

As discussed in Section 1, exiting calibration measures cannot distinguish between overconfidence and underconfidence. In the NTCIR-19 R2C2 task, we leverage HMR and its component measures called *rewards* for suppressing overconfidence and underconfidence (R_O, R_U). For a given system, let I^- and I^+ denote the sets of questions for which the system's answers are considered incorrect and correct, respectively ($|I^-| + |I^+| = N$). Let $p(i)$ denote the system's confidence for Question i. Then, for each $i \in I^-$ (the system is incorrect), $p(i)$ should be as close to 0 as possible; whereas for $i \in I^+$ (the system is correct), $p(i)$ should be as close to 1 as possible. Hence, the rewards (R_O, R_U are defined as follows [6].

$$O = \sum_{i \in I^-} p(i) \ , \ \ U = \sum_{i \in I^+} (1 - p(i)) \ , \tag{1}$$

$$R_O = \begin{cases} 1 & \text{if } I^- = \emptyset \ , \\ 1 - O/|I^-| & \text{otherwise} \ . \end{cases} \tag{2}$$

$$R_U = \begin{cases} 1 & \text{if } I^+ = \emptyset \ , \\ 1 - U/|I^+| & \text{otherwise} \ . \end{cases} \tag{3}$$

Note, for example, that when $I^- = \emptyset$ (i.e., all N system responses are correct), there is no way for the system to be overconfident for any of the instances and therefore $R_O = 1$ (i.e., perfection).

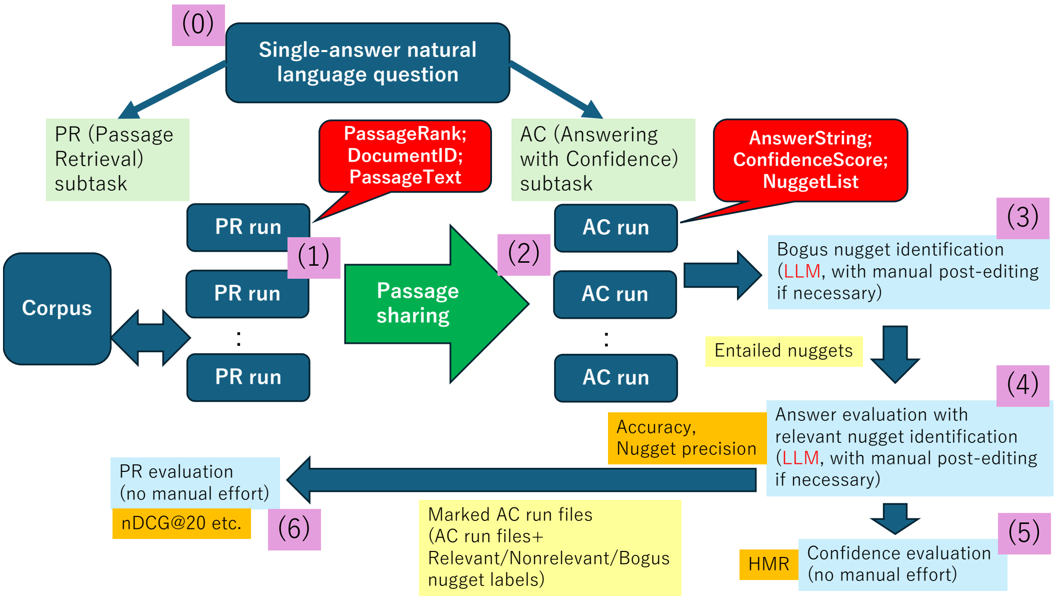

(0)
Single-answer natural language question
PR (Passage Retrieval) subtask
PassageRank; DocumentID; PassageText
AC (Answering with Confidence) subtask
AnswerString; ConfidenceScore; NuggetList
Corpus
PR run
PR run
:
PR run
(1)
Passage sharing
(2)
AC run
AC run
:
AC run
(3)
Bogus nugget identification (LLM, with manual post-editing if necessary)
Entailed nuggets
(4)
Answer evaluation with relevant nugget identification (LLM, with manual post-editing if necessary)
Accuracy, Nugget precision
PR evaluation (no manual effort)
nDCG@20 etc.
(6)
Marked AC run files (AC run files+ Relevant/Nonrelevant/Bogus nugget labels)
HMR
Confidence evaluation (no manual effort)
(5)

Figure 2: Overview of the NTCIR-19 R2C2 task.

Table 1
Timeline for the NTCIR-19 R2C2 Task. We will release the corpus in Nov 2025 and then conduct a dry run immediately after.

Date	Event
Mar 13, 2026	Topics released
Apr 17, 2026	PR-runs due; directly shared to participants; task registrations due
May 15, 2026	AC-runs due
Aug 1, 2026	evaluation results and draft overview released
Sep 1, 2026	draft participant papers due
Nov 1, 2026	all camera ready papers due
Dec 8-10 2026	NTCIR-19 conference (Tokyo)

As we want systems to balance the above two rather than to sacrifice one for the sake of the other, HMR employs the Harmonic Mean [65]:

$$HMR = \begin{cases} 0 & \text{if } R_O = R_U = 0\,, \\ 2\,R_O\,R_U/(R_O + R_U) & \text{otherwise}\,. \end{cases} \tag{4}$$

3. The R2C2 Task

3.1. Overview

Figure 2 provides a visual overview of the NTCIR-19 R2C2 task. Table 1 shows the timeline. Below, we describe the seven steps indicated in the figure.

(0) We first make the test question set publicly available so that *Passage Retrieval* (PR) and *Answering with Confidence* (AC) subtask participants can get to work. Each question is a movie-related question that requires a single answer—a phrase, sentence, quotation from a passage, etc. For example: "*Name the actor who starred in a movie called The Manchurian Candidate and later played Captain America.*" (The correct answer is *Anthony Mackie.*[3])

[3] See https://www.imdb.com/title/tt0368008/fullcredits/ (The Manchurian Candidate (2004)) and https://www.imdb.com/title/tt14513804/ (Captain America: Brave New World) .

(1) We collect PR runs from the participants, and make them available to all AC subtask participants, so that they can utilise any of the PR runs—even those from a different team. This was inspired by the Reliable Information Access workshop held in 2003, where pseudo-relevance feedback was studied by exchanging initial search results across workshop participants [66]. Thus, potentially, a team that is particularly good at the "R" part of RAG and another team that is particularly good at the "G" part of RAG can synergise. A PR run contains, for each test question, a ranked list of generated passages, each with a source document ID from the corpus. This is actually a "passage generation and ranking" task: the PR subtask participants can either (a) Utilise baseline passages (with source document IDs) that the organisers provide "as is" or (b) Generate their own passages from the corpus provided by the organisers. More details will be given in Section 3.4.

(2) We collect AC runs from the participants. As detailed in Section 3.5, an AC run contains, for each question, an answer string with a confidence score, as well as a list of nuggets extracted from the passages of any of the PR runs. Each nugget must be accompanied by a *PassageKey*, comprised of a PR run file name and the rank of a passage within that run.

(3) We conduct *bogus nugget identification* using an LLM (with manual post-editing of necessary). Given a passage text and a nugget, we say that the nugget is *bogus* iff the passage text does not *entail* the nugget. For example, if a passage text about The Manchurian Candidate (1962)[4] contains "*stars Frank Sinatra Laurence Harvey Janet Leigh*" (where Frank Sinatra, Laurence Harvey, and Janet Leigh are actor names) and a nugget that cites this passage (through a PassageKey) claims that "The Manchurian Candidate starred **Harvey Janet**," then that nugget would be considered bogus. Thus, only nuggets that are entailed by the cited passages are passed on to the next evaluation step.

(4) We employ another LLM to conduct *answer evaluation with relevant nugget identification* for each answer in each AC run file. More specifically, we will provide the answer string and the list of entailed nuggets to an LLM, and ask it whether the answer is correct assuming that the nuggets are factually correct, and if so, which of the nuggets helped to deduce that answer. Nuggets that helped are considered *relevant*; those that did not help (although they are entailed by the passages) are considered *nonrelevant*. For example, an entailed nugget about Frank Sinatra from a passage from The Manchurian Candidate (1962) movie is unrelated to the correct answer *Anthony Mackie* (who starred in a 2004 movie of the same name) and therefore should be considered nonrelevant in this context. Thus, for the entire test question set, we can compute the *Accuracy* (i.e., proportion of correct answers); for each question, we can compute *Nugget Precision* (i.e., proportion of relevant nuggets among all nuggets returned (including bogus ones)). We will also compute Mean Nugget Precision over the question set.

(5) Having determined whether each answer is correct, we examine whether the confidence scores that accompany the answers are appropriate using the HMR measure described in Section 2.4. We thus evaluate modesty: the ability to suppress overconfidence *and* underconfidence.

(6) Finally, we evaluate the PR runs without conducting passage relevant assessments. We will achieve this by deducing the relevance grades of the passages based on how the passages contributed to finding the correct answers in the AC subtask. More specifically, a passage that provided n relevant nuggets across all AC runs will be given a relevance grade of n. We will compute ranked retrieval evaluation measures such as nDCG, and then compute mean scores over the question set.

Note that we employ a pipeline approach in Steps (3) and (4) when relying on LLMs. This strategy is based on our view that, if LLMs' behaviours are not altogether predictable, then we should try to reduce the number of confounding uncertainty factors, by giving them simple and controllable subtasks instead of a complex task. That is, divide and conquer. This point probably deserves some experimental verification in our future work.

[4] https://www.imdb.com/title/tt0056218/

3.2. Questions

In this first round of the R2C2 task, we will deal with natural language questions that require only one answer (a phrase, sentence, quotation from a passage, etc.). Each question will be something to do with movies, and will be deliberately designed to test RAG's ability to retrieve relevant passages and to utilise them to derive the right answer. Yang et al. [67] describe a question taxonomy for benchmarking RAG systems, which contains the following categories: *Simple, Simple w. Condition, Set, Comparison, Aggregation, Multi-hop, Post-processing-heavy, and False Premise.* From these, we plan to include the following in our test question set; the definitions and examples are copied verbatim from Yang et al. [67, Table 2].

Simple "Questions asking for simple facts that are unlikely to change over time, such as the birth date of a person and the authors of a book."

Simple w. Condition "Questions asking for simple facts with some given conditions, such as stock prices on a certain date and a director's recent movies in a certain genre."

Comparison "Questions that compare two entities (e.g., "*who started performing earlier, Adele or Ed Sheeran?*")."

Aggregation Questions that require aggregation of retrieval results to answer (e.g., "*how many Oscar awards did Meryl Streep win?*").

Multi-hop "Questions that require chaining multiple pieces of information to compose the answer (e.g., "*who acted in Ang Lee's latest movie?*")."

Post-processing-heavy "Questions that need reasoning or processing of the retrieved information to obtain the answer (e.g., "*how many days did Thurgood Marshall serve as a Supreme Court justice?*")."

We rule out the **Set** category as we are dealing with single-answer questions rather than those that require a set of answers. We also rule out the **False Premise** category (i.e., questions that include a false assumption), because the correct response for this category would be "*I don't have an answer to your question. Your question is based on a false premise [...]*" and evaluating it would require special handling.[5] Our example question, "*Name the actor who starred in a movie called The Manchurian Candidate and later played Captain America.*" is a Multi-hop question, as it probably requires looking up the cast for each of the movies and then take the intersection.[6]

In the dry run phase, we will only use about five questions, just to check that our evaluation protocol is feasible. We will then compute a two-way ANOVA residual variance as a rough estimate of the common population variance (under the homoskedasticity assumption) of some of the aforementioned evaluation measures (e.g., nDCG) from the question-by-run score matrices, using the procedure described in Sakai [68, 69]. This will enable us to consider appropriate sample sizes for the test questions in advance. Moreover, since the above score matrices may be too small for estimating the population variance„ we will consider *pooling* estimated variances of nDCG at 20 etc. from several topic-by-run score matrices from past NTCIR IR task results to enhance the estimation accuracy. See Sakai [68, Section 6.7] for details.

One of the reviewers of this paper suggested leveraging existing QA datasets. However, our view is that such an approach is vulnerable to *contamination*: the participating LLMs may have already seen these questions and we may overestimate their effectiveness. Hence we prefer to create our own questions. Another reviewer comment regarding question creation was: "*Why movies*?" The main reason is that, based on our experience with running the NTCIR FairWeb tasks [70, 18] where we also handled movie-related topics, movies and movie-related entities provide excellent materials for

[5] One reviewer maintained that the False Premise category should not be skipped. See also Section 4.3 for our view on this category.

[6] unless there exists a single passage that directly answers this question–but that is highly unlikely as this is a very nerdy question composed by the first author of this paper.

constructing comparison, aggregation, multi-hop, and post-processing-heavy questions and there is plenty of publicly available information sources for answering these questions; we will discuss the corpus in the next section. Another reason is that some of us organisers consider ourselves to be nerds on some movies, such as Star Wars: we felt confident about creating topics from the viewpoint of an "expert."

3.3. Corpus

RAG systems usually work with a domain-specific corpus. We have therefore constructed a corpus containing movie related documents. To do this, we crawled Wikepedia, as well as "Wookieepedia"[7] which specialises in Star Wars-related contents. When creating each question, we checked that it can be answered correctly using either Wikepedia or Wookiepedia.

3.3.1. Wikipedia

To construct a Wikipedia-based corpus covering almost all movie-related entities, we systematically extracted documents related to movies, directors, actors, movie-related organizations, awards, festivals, etc. from the publicly available dumps of Wikidata and Wikipedia.

Wikidata[8] provides structured semantic relations, which enable reliable identification of entities in the movie domain, while Wikipedia offers natural language texts associated with these entities.

We first identified movie-related entities from the official Wikidata dump. [9] The dump, originally released by Wikidata on 26 September, 2025, was downloaded on 13 October, 2025. For every entity in the dump, we investigated its structured attributed (claims) to determine whether it is movie-related.

Entities categorized as films, film series, film festivals, film awards, film genres, or film-related organizations (e.g., production companies and studios) were retained. We also included entries describing film industry professionals such as directors, actors, and producers. Additionally, we extracted entities that represent lists of movies, characters, movie-related personnel, relevant events and awards. For each qualified entity, we recorded its ID and title for later text retrieval. In total, we identified 338,761 movie-related entities.

Using the titles obtained from Wikidata, we downloaded the original wikitext code of each corresponding article through the MediaWiki API.[10] The wikitext code preserves the complete structure of the article, including filmographies, infoboxes, tables, and multi-column lists. The downloading process was performed in small batches (ten titles per request) with automatic retry and rate-limiting to respect the API usage guidelines. Each article was saved as an individual document in a certain format with a short metadata header:

```
File name: Q25540924__Star_Wars_prequel_trilogy.wikitext
# QID: Q25540924
# TITLE: Star_Wars_prequel_trilogy
# LANG: en
```

From the identified 338,761 entities, we successfully downloaded a total of 332,378 documents.

3.3.2. Wookieepedia

Wookieepedia is the Star Wars community encyclopedia hosted on Fandom[11]. We first downloaded the full-site dump provided by Wookieepedia [12], which contains all current pages of Wookieepedia in

[7] https://starwars.fandom.com/wiki/Main_Page

[8] https://www.wikidata.org/

[9] https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.json.bz2, approximately 100GB compressed, 500GB uncompressed

[10] https://www.mediawiki.org/wiki/API

[11] https://www.fandom.com/

[12] https://s3.amazonaws.com/wikia_xml_dumps/s/st/starwars_pages_current.xml.7z The download link can be found at https://starwars.fandom.com/wiki/Special:Statistics.

XML format. The dump used for our corpus construction was downloaded on 21 September, 2025. This version had been released by Wookieepedia on 16 September, 2025.

After obtaining the dump, we extract the articles in wikitext code, and save them in a similar way:

```
File name: 17__Star_Wars_Episode_II_Attack_of_the_Clones.wikitext
# ID: 17
# TITLE: Star_Wars_Episode_II_Attack_of_the_Clones
# LANG: en
```

As a result, a total of 219,285 documents were extracted from the Wookieepedia dump.

3.3.3. Baseline passages

As some PR subtask participants may not want to generate their own passages, we are providing baseline passages to them as well, where each passage is accompanied by a source document ID. Hence, PR subtask participants may choose to just index the baseline passages and then rank them for each question.

We prepared the baseline passages through the following steps. The raw wikitext code was cleaned using mwparserfromhell[13] with regular-expression fallbacks, removing templates, HTML tags, and comments, while converting internal and external links to plain text. Table blocks were converted into inline Markdown tables by wikitextparser,[14] preserving the textual content of each cell. The resulting plain text was then chunked into overlapping passages using a fixed window of 400 words with a 15% overlap. Each passage (chunk) was stored in JSON format, containing the passage text along with its word count, overlap ratio with the previous passage, passage ID, and the corresponding source document ID and title. All passages belonging to the same document were saved together in a compressed JSONL file (one file per document), and each file was named after the source document ID. As a result, we obtained 1,101,442 passages in total, including 848,803 from Wikipedia and 252,639 from Wookieepedia documents.

3.4. Passage Ranking Subtask

This section briefly describes the specification of the PR subtask.

As discussed earlier, the organisers will provide a corpus that covers movie-related topics to participants. We will also provide baseline passages extracted from each document. Thus, PR subtask participants may either extract passages from each document using their own method, or rely on our baseline passages. Recall that this is not a *passage ID* ranking task: we require systems to rank document IDs with passage texts.

The input to a PR system is a natural language question: recall the Captain America example mentioned earlier.

The output required is a ranked list of passages, containing no more than 20 passages. More specifically, a ranked list in a PR run file should contain the following four fields (separated by a semicolon).

QuestionID Question ID given in the test question file;

PassageRank An integer in the [1-20] range, representing the rank of the passage (ties are not allowed);

DocID Document ID given in the corpus;

PassageText A plain-text passage that represents one consecutive part of the above document (no line breaks allowed).

[13] https://github.com/earwig/mwparserfromhell

[14] https://github.com/5j9/wikitextparser

As we have briefly described in Section 3.1, we will derive the relevance grade of each passage from the results of the AC task. Within the context of a particular question, a combination of *PRrunName* (PR run file name) and a *PassageRank*, forms a *key* for a passage, in the database sense; we refer to this as *PassageKey*. A PassageKey that provided n relevant nuggets across all AC runs (for a particular question) will be given a relevance grade of n. Thus, we will create a qrels file, where each line consists of a QuestionID, a PassageKey, and a RelevanceGrade.

From each PR run, we will create a "res" (result) file [69], where each line will contain a PassageKey: that is, each line will include a RunName and a PassageRank. Thus, we can use the NTCIREVAL evaluation tool [69] to compute IR evaluation measures just like many previous IR tasks did at NTCIR (e.g., [18]).

3.5. Answering with Confidence Subtask

This section briefly describes the specification of the AC subtask.

The input to an AC system is the same natural language question used in the PR subtask, as well as any of the PR runs chosen by the AC subtask participating teams.

An AC run file is required to contain the following for each question.

AnswerString A word, phrase, sentence, quotation from a passage, etc. Recall that sets and lists etc. are not expected as answers.

ConfidenceScore An integer in the 0-100 range that represents the system's confidence about the above response.

NuggetList A list containing up to 10 lines. Each line represents a nugget extracted by the AC run, and is composed of a *NuggetNum* (a positive integer in the 1-10 range that represents a nugget), a *PassageKey*, and a *Nugget* (a text generated based on the passage). We call such line a *NuggetRecord*.

To evaluate the AC runs, we first conduct bogus nugget identification using an LLM, as described in Step (3) of Section 3.1. The input to our bonus nugget remover are a passage-nugget pair; the output is a YES or NO. If the passage entails the nugget, the output label should be YES, otherwise, it should be NO (i.e., the nugget is bogus). To indicate bogus nuggets, we will add "B" as a prefix to the nugget record in the AC run file. To facilitate manual post-editing, we plan to make LLM output a reason for the YES/NO decision as well.

In the answer evaluation with relevant nugget identification phase, bogus nuggets are ignored; that is, only entailed nuggets are considered. The input to the LLM-based answer evaluator are the *test question*, *AnswerString*, and the entailed nuggets. The answer evaluator is asked to answer two questions.

Q1 Assuming that the nuggets are factually correct, is the returned answer correct? YES or NO?

Q2 Which of the nuggets help to figure out that answer? Mark those who help as *relevant*, those who do not as *nonrelevant*.

In the AC run file, we will add "R" as a prefix to relevant nuggets, and "N" as a prefix to nonrelevant nuggets. The modified AC run file with the prefixes (B, R, or N) is called a *marked* AC run file. Again, to facilitate manual post-editing, we plan to make LLM output a reason for the YES/NO decision, as well as why it considers each nugget to be (non)relevant.

Based on marked AC run files, we can compute the accuracy, mean nugget precision scores, as well as HMR scores.

4. Potential Challenges and Future Work

As the new R2C2 task is at the preparation stage at the time of writing, this section discusses potential challenges of this task.

4.1. PR and AC: Best Of Both Worlds?

One key feature of this task is passage sharing, in the hope that the best PR approaches and the best AC approaches can be combined, even across different teams. However, in order to thus enjoy the best of both worlds, the following must happen.

- We will need several good participating teams in this task. Consider an extreme case where there is only one participating team: then there would be no option to utilise a PR run from another team.
- We will need decent PR runs. If the "R" step of RAG systems is totally unsuccessful, then the "G" step will not be able to do anything.
- We will need decent AC runs. If the "G" step of RAG systems is completely unable to derive good answers even though the retrieved passages are *actually* relevant, note that our PR evaluation will not work: the truly relevant documents will be considered nonrelevant unless they bear at least one relevant nugget in the AC evaluation.

To avoid worst case scenarios, the organisers will try to provide decent PR and AC runs as baselines. In particular, we will consider providing an *oracle* PR run that contain manually identified relevant passages: will these passages actually be considered relevant in our automatic PR evaluation step? Moreover, we will consider LLM-based or LLM-assisted relevant assessments (e.g., [47, 43, 71, 45]) of passages to check if our automatic PR evaluation scheme actually misses truly relevant documents.

4.2. Can We Trust LLMs as Judges?

As we have discussed in Section 3.1, we adopt a divide and conquer approach and give LLMs relatively simple tasks, namely, bogus nugget identification and answer evaluation with relevant nugget identification. However, we cannot guarantee 100% accuracy and *some* manual post-editing may be required. The question is, how much? Whatever the answer is, we need to make sure that the LLM-based evaluation aligns with what humans would do, to avoid the aforementioned problems such as *narcissism* (See Section 2.3).

4.3. Which Question Types? How Do We Evaluate "Not Answering"?

Section 3.2 discussed the question types that we will cover in this round of the R2C2 task. How well do current RAG systems do for each question type? Should we include some Set and False Premise questions in the next round? We have mentioned in Section 1 that RAG systems can in principle utilise low confidence scores (if they align well with accuracy) to determine whether they should respond with "*I don't know*" (IDK). Handling the evaluation of IDK [2] along with responses such as "*I can't answer that question because it is based on a false premise*" may be worth tackling. Since there are many possible ways to tell the user that the system cannot answer (for some reason), one possible approach to evaluating systems with questions would be to employ an LLM to identify a *dialogue act* of the RAG system response (which is referred to as Type-O nuggets in Sakai [7]). For example, both of the above responses could be treated as a *refusal to answer* dialogue act and then perhaps the responses can be evaluated accordingly.

Another point that deserves discussion is that RAG benchmark questions tend to be highly artificial: they are designed to test the deduction capabilities of RAG systems, and do not necessarily reflect real user needs. How do we address this issue? Should we?

4.4. Modesty: Are RAG Systems Almost Always Overconfident?

The advantage of our modesty evaluation framework using the HMR measure over traditional calibration evaluation is that we can quantify overconfidence and underconfidence separately. We anticipate that RAG systems are often overconfident but that they can be underconfident at times. Through our evaluation effort, how can we make RAG systems suppress both overconfidence and underconfidence at the same time rather than suppressing one at the expense of tolerating the other? For what question types and what kinds of systems will we witness overconfidence or underconfidence? We hope that the official R2C2 evaluation results will shed some light on these points.

4.5. Multiple Answers? Multiple Confidence Scores?

This round of the R2C2 task deals with single-answer questions only: one answer and one confidence score per question. If we want to extend our evaluation framework to more complex questions, there would be several choices, e.g., multiple answers where each answer is accompanied by a confidence score, multiple answers with a single overall confidence score, and so on. Moreover, multiple answers could be a set, a ranked list, or even a mixture of these if the question asks for multiple types of information in a single turn. How can we handle these situations? Should we?

5. Conclusions

We described the ongoing NTCIR-19 R2C2 task, which requires RAG systems to return an appropriate confidence score together with the answer to each question, and discussed the potential risks and challenges. Our key idea is to evaluate the *modesty* of RAG systems, that is, whether the confidence scores are appropriately high when the answers are correct, and whether they are appropriately low when the answers are incorrect. “Modest” RAG systems with an appropriate confidence threshold will know exactly when to return IDK. At SIGIR-AP 2025, we hope to discuss with the BREV-RAG workshop participants how to move forward with the task, as well as how our community should move forward towards evaluating RAG systems from various axes.

Acknowledgement

We thank the BREV-RAG reviewers for their insightful and constructive comments.

Declaration on Generative AI

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

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: http://dx.doi.org/10.1145/3571730. doi:10.1145/3571730.

[2] A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, Why language models hallucinate, 2025. URL: https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf.

[3] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment, 2023. arXiv:2308.05374.

[4] D. Yang, L. Zeng, J. Rao, Y. Zhang, Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 1305–1315. URL: https://doi.org/10.1145/3726302.3730018. doi:10.1145/3726302.3730018.

[5] M. Pakdaman Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, Proceedings of the AAAI Conference on Artificial Intelligence 29 (2015). URL: https://ojs.aaai.org/index.php/AAAI/article/view/9602.

[6] 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.

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

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[9] F. Petroni, F. Siciliano, F. Silvestri, G. Trappolini, IR-RAG @ SIGIR24: Information retrieval's role in RAG systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3036–3039. URL: https://doi.org/10.1145/3626772.3657984. doi:10.1145/3626772.3657984.

[10] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, N. McAleese, Teaching language models to support answers with verified quotes, 2022. URL: https://arxiv.org/abs/2203.11147. arXiv:2203.11147.

[11] J. Wallat, M. Heuss, M. d. Rijke, A. Anand, Correctness is not faithfulness in retrieval augmented generation attributions, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 22–32. URL: https://doi.org/10.1145/3731120.3744592. doi:10.1145/3731120.3744592.

[12] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, J. Kaplan, A general language assistant as a laboratory for alignment, 2021. URL: https://arxiv.org/abs/2112.00861. arXiv:2112.00861.

[13] E. Dinan, G. Abercrombie, A. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, V. Rieser, SafetyKit: First aid for measuring safety in open-domain conversational systems, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4113–4133. URL: https://aclanthology.org/2022.acl-long.284/. doi:10.18653/v1/2022.acl-long.284.

[14] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, 2023. URL: https://arxiv.org/abs/2305.17926. arXiv:2305.17926.

[15] B. Murugadoss, C. Poelitz, I. Drosos, V. Le, N. McKenna, C. S. Negreanu, C. Parnin, A. Sarkar, Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions, 2024. URL: https://arxiv.org/abs/2408.08781. arXiv:2408.08781.

[16] N. Chen, Q. Dai, X. Dong, P. Wang, Q. Jia, Z. Du, Z. Dong, X.-M. Wu, Evaluating conversational recommender systems via large language models: A user-centric framework, 2025. URL: https://arxiv.org/abs/2501.09493. arXiv:2501.09493.

[17] M. D. Ekstrand, G. McDonald, A. Raj, I. Johnson, Overview of the TREC 2022 fair ranking track, in: Proceedings of TREC 2022, 2023. URL: https://trec.nist.gov/pubs/trec31/papers/Overview_fair.pdf.

[18] 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.

[19] T. E. Kim, F. Diaz, Towards fair rag: On the impact of fair ranking in retrieval-augmented generation, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 33–43. URL: https://doi.org/10.1145/3731120.3744599. doi:10.1145/3731120.3744599.

[20] M. Mousavian, Z. Abbasiantaeb, M. Aliannejadi, F. Crestani, Towards fair rankings: Leveraging LLMs for gender bias detection and measurement, in: Proceedings of the 2025 International

ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 56–66. URL: https://doi.org/10.1145/3731120.3744620. doi:10.1145/3731120.3744620.

[21] T. Sakai, Fairness-based evaluation of conversational search: A pilot study, in: Proceedings of EVIA 2023, 2023. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings17/pdf/evia/01-EVIA2023-EVIA-SakaiT.pdf.

[22] 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.

[23] N. Arabzadeh, C. L. A. Clarke, A comparison of methods for evaluating generative IR, 2024. URL: https://arxiv.org/abs/2404.04044. arXiv:2404.04044.

[24] L. Gienapp, H. Scells, N. Deckers, J. Bevendorff, S. Wang, J. Kiesel, S. Syed, M. Fröbe, G. Zuccon, B. Stein, M. Hagen, M. Potthast, Evaluating generative ad hoc information retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1916–1929. URL: https://doi.org/10.1145/3626772.3657849. doi:10.1145/3626772.3657849.

[25] E. M. Voorhees, Overview of the TREC 2003 question answering track, in: Proceedings of TREC 2003, 2004. URL: https://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf.

[26] A. Nenkova, R. Passonneau, K. McKeown, The pyramid method: Incorporating human content selection variation in summarization evaluation, ACM Trans. Speech Lang. Process. 4 (2007) 4–es. URL: https://doi.org/10.1145/1233912.1233913. doi:10.1145/1233912.1233913.

[27] H. T. Dang, J. Lin, Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors, in: A. Zaenen, A. van den Bosch (Eds.), Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 768–775. URL: https://aclanthology.org/P07-1097/.

[28] T. Mitamura, H. Shima, T. Sakai, N. Kando, T. Mori, K. Takeda, C.-Y. Lin, R. Song, C.-J. Lin, C.-W. Lee, Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access, in: Proceedings of NTCIR-8, 2010, pp. 15–24. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-CCLQA-MitamuraT.pdf.

[29] T. Sakai, M. P. Kato, Y.-I. Song, Click the search button and be happy: evaluating direct and immediate information access, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 621–630. URL: https://doi.org/10.1145/2063576.2063669. doi:10.1145/2063576.2063669.

[30] T. Sakai, Z. Dou, Summaries, ranked retrieval and sessions: a unified framework for information access evaluation, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 473–482. URL: https://doi.org/10.1145/2484028.2484031. doi:10.1145/2484028.2484031.

[31] M. Kato, T. Sakai, T. Yamamoto, V. Pavlu, H. Morita, S. Fujita, Overview of the NTCIR-12 MobileClick-2 task, in: Proceedings of NTCIR-12, 2016, pp. 104–114. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/OVERVIEW/01-NTCIR12-OV-MOBILECLICK-KatoM.pdf.

[32] Z. Zeng, C. Luo, L. Shang, H. Li, T. Sakai, Test collections and measures for evaluating customer-helpdesk dialogues, in: Proceedings of EVIA 2017, 2017, pp. 1–9. URL: https://ceur-ws.org/Vol-2008/paper_1.pdf.

[33] M. Alaofi, N. Arabzadeh, C. L. A. Clarke, M. Sanderson, Generative Information Retrieval Evaluation, Springer Nature Switzerland, 2024, p. 135–159. URL: http://dx.doi.org/10.1007/978-3-031-73147-1_6. doi:10.1007/978-3-031-73147-1_6.

[34] J. Mayfield, E. Yang, D. Lawrie, S. MacAvaney, P. McNamee, D. W. Oard, L. Soldaini, I. Soboroff, O. Weller, E. Kayi, K. Sanders, M. Mason, N. Hibbler, On the evaluation of machine-generated

reports, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1904–1915. URL: https://doi.org/10.1145/3626772.3657846. doi:10.1145/3626772.3657846.

[35] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, J. Lin, The great nugget recall: Automating fact extraction and rag evaluation with large language models, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 180–190. URL: https://doi.org/10.1145/3726302.3730090. doi:10.1145/3726302.3730090.

[36] Z. Abbasiantaeb, S. Lupart, L. Azzopardi, J. Dalton, M. Aliannejadi, Conversational gold: Evaluating personalized conversational search system using gold nuggets, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 3455–3465. URL: https://doi.org/10.1145/3726302.3730316. doi:10.1145/3726302.3730316.

[37] E. Yang, D. Lawrie, H. Dang, I. Soboroff, J. Mayfield, Nugget-based annotation protocol and tool for evaluating long-form retrieval-augmented generation, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 3999–4003. URL: https://doi.org/10.1145/3726302.3730156. doi:10.1145/3726302.3730156.

[38] C. Samarinas, A. Krubner, A. Salemi, Y. Kim, H. Zamani, Beyond factual accuracy: Evaluating coverage of diverse factual information in long-form text generation, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 13468–13482. URL: https://aclanthology.org/2025.findings-acl.693/. doi:10.18653/v1/2025.findings-acl.693.

[39] N. Thakur, J. Lin, S. Havens, M. Carbin, O. Khattab, A. Drozdov, Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents, 2025. URL: https://arxiv.org/abs/2504.13128. arXiv:2504.13128.

[40] H. Hashemi, J. Eisner, C. Rosset, B. Van Durme, C. Kedzie, LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13806–13834. URL: https://aclanthology.org/2024.acl-long.745/. doi:10.18653/v1/2024.acl-long.745.

[41] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, S. Ma, Leveraging large language models for NLG evaluation: Advances and challenges, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16028–16045. URL: https://aclanthology.org/2024.emnlp-main.896/. doi:10.18653/v1/2024.emnlp-main.896.

[42] J.-L. Peng, S. Cheng, E. Diau, Y.-Y. Shih, P.-H. Chen, Y.-T. Lin, Y.-N. Chen, A survey of useful LLM evaluation, 2024. URL: https://arxiv.org/abs/2406.00936. arXiv:2406.00936.

[43] C. Clarke, L. Dietz, LLM-based relevance assessment still can't replace human relevance assessment, in: Proceedings of EVIA 2025, 2025, pp. 1–5. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings18/pdf/evia/01-EVIA2025-EVIA-ClarkeC.pdf.

[44] I. Soboroff, Don't use LLMs to make relevance judgments, Information Retrieval Research 1 (2025) 29–46. URL: https://irrj.org/article/view/19625. doi:10.54195/irrj.19625.

[45] R. Takehi, E. M. Voorhees, T. Sakai, I. Soboroff, LLM-assisted relevance assessments: When should we ask LLMs for help?, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 95–105. URL: https://doi.org/10.1145/3726302.3729916. doi:10.1145/3726302.3729916.

[46] S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, J. Lin, A large-scale study of relevance assessments with large language models: An initial look, 2024. URL:

https://arxiv.org/abs/2411.08275. arXiv:2411.08275.

[47] M. Alaofi, P. Thomas, F. Scholer, M. Sanderson, LLMs can be fooled into labelling a document as relevant: best café near me; this paper is perfectly relevant, 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, Association for Computing Machinery, New York, NY, USA, 2024, p. 32–41. URL: https://doi.org/10.1145/3673791.3698431. doi:10.1145/3673791.3698431.

[48] A. Panickssery, S. R. Bowman, S. Feng, LLM evaluators recognize and favor their own generations, 2024. URL: https://arxiv.org/abs/2404.13076. arXiv:2404.13076.

[49] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, 2017. arXiv:1706.04599.

[50] Z. Jiang, J. Araki, H. Ding, G. Neubig, How can we know when language models know? on the calibration of language models for question answering, Transactions of the Association for Computational Linguistics 9 (2021) 962–977. URL: https://aclanthology.org/2021.tacl-1.57. doi:10.1162/tacl_a_00407.

[51] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, M. Lucic, Revisiting the calibration of modern neural networks, 2021. URL: https://arxiv.org/abs/2106.07998. arXiv:2106.07998.

[52] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2023. arXiv:2211.09110.

[53] S. Lin, J. Hilton, O. Evans, Teaching models to express their uncertainty in words, 2022. arXiv:2205.14334.

[54] K. Nguyen, B. O'Connor, Posterior calibration and exploratory analysis for natural language processing models, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1587–1598. URL: https://aclanthology.org/D15-1182. doi:10.18653/v1/D15-1182.

[55] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang, D. Tran, Measuring calibration in deep learning, 2020. arXiv:1904.01685.

[56] G. Portillo Wightman, A. Delucia, M. Dredze, Strength in numbers: Estimating confidence of large language models by prompt agreement, in: A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (Eds.), Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 326–362. URL: https://aclanthology.org/2023.trustnlp-1.28. doi:10.18653/v1/2023.trustnlp-1.28.

[57] W. Tam, X. Liu, K. Ji, L. Xue, J. Liu, T. Li, Y. Dong, J. Tang, Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 13117–13130. URL: https://aclanthology.org/2023.findings-emnlp.874. doi:10.18653/v1/2023.findings-emnlp.874.

[58] H. Wang, Z. Zhang, M. Hu, Q. Wang, L. Chen, Y. Bian, B. Wu, RECAL: Sample-relation guided confidence calibration over tabular data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7246–7257. URL: https://aclanthology.org/2023.findings-emnlp.482. doi:10.18653/v1/2023.findings-emnlp.482.

[59] P. Zablotskaia, D. Phan, J. Maynez, S. Narayan, J. Ren, J. Liu, On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association

for Computational Linguistics, Singapore, 2023, pp. 2980–2992. URL: https://aclanthology.org/2023.findings-emnlp.197. doi:10.18653/v1/2023.findings-emnlp.197.

[60] C. Zhu, B. Xu, Q. Wang, Y. Zhang, Z. Mao, On the calibration of large language models and alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9778–9795. URL: https://aclanthology.org/2023.findings-emnlp.654. doi:10.18653/v1/2023.findings-emnlp.654.

[61] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1950) 1 – 3. URL: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml. doi:https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[62] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. arXiv:1906.02530.

[63] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, C. Manning, Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5433–5442. URL: https://aclanthology.org/2023.emnlp-main.330. doi:10.18653/v1/2023.emnlp-main.330.

[64] K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, R. Hartley, Calibration of neural networks using splines, 2021. arXiv:2006.12800.

[65] T. Sakai, Metrics, Statistics, Tests, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 116–163. URL: https://doi.org/10.1007/978-3-642-54798-0_6. doi:10.1007/978-3-642-54798-0_6.

[66] D. Harman, C. Buckley, The NRRC reliable information access (RIA) workshop, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 528–529. URL: https://doi.org/10.1145/1008992.1009104. doi:10.1145/1008992.1009104.

[67] X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, L. Kong, B. Moran, J. Wang, Y. E. Xu, A. Yan, C. Yang, E. Yuan, H. Zha, N. Tang, L. Chen, N. Scheffer, Y. Liu, N. Shah, R. Wanga, A. Kumar, W. tau Yih, X. L. Dong, CRAG – comprehensive RAG benchmark, 2024. URL: https://arxiv.org/abs/2406.04744. arXiv:2406.04744.

[68] T. Sakai, Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, Springer, 2018.

[69] T. Sakai, How to Run an Evaluation Task, Springer International Publishing, Cham, 2019, pp. 71–102. URL: https://doi.org/10.1007/978-3-030-22948-1_3. doi:10.1007/978-3-030-22948-1_3.

[70] 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.

[71] T. Sakai, K. M. Rain, R. Takehi, S. Tao, Y.-I. Song, Open-source LLM-based relevance assessment vs. highly reliable manual relevance assessment: A case study, in: Proceedings of CIKM 2025, 2025, p. to appear.