Evaluation Validity in Information Retrieval

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

Nick Craswell
Microsoft
Redmond, United States
nickcr@microsoft.com

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Seth Spielman
Microsoft
Boulder, United States
sethspielman@microsoft.com

Robert Sim
Microsoft
Redmond, United States
rsim@microsoft.com

Ryen W. White
Microsoft
Redmond, United States
ryenw@microsoft.com

Abstract

Information retrieval has long relied on evaluations that measure system performance. Improvements on standard evaluation protocols are interpreted as progress in system effectiveness, on the understanding that improved metrics indicate a better experience. However, most evaluations are a drastic abstraction and simplification of that experience. It is reasonable to inquire after the *validity* of our evaluations, or the degree to which they do in fact represent phenomena we care about. *If a metric improves, can we be sure there is a corresponding improvement in real-world effectiveness?*

We discuss practical ways to discuss, measure, and improve the validity of evaluations in a range of settings. By considering validity, we can make better choices in evaluation protocols; we have a chance to make progress if and when evaluating and retrieving collapse into each other entirely, e.g., with LLM-as-judge; and we can optimise towards systems that people actually want.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**; • **General and reference** → **Evaluation**; *Metrics*; *Experimentation*; • **Human-centered computing** → *Natural language interfaces.*

Keywords

Validity; metrics; retrieval-augmented generation

ACM Reference Format:
Paul Thomas, Nick Craswell, Mark Sanderson, Seth Spielman, Robert Sim, and Ryen W. White. 2026. Evaluation Validity in Information Retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3805712.3808538

1 Introduction

Evaluation has long been central to Information Retrieval (IR), and researchers and practitioners have developed a broad set of evaluation techniques [16, 27, 47, 54, 89, 125]. Each purports to tell us something about how well a system achieves its designers' aims: improvements on an evaluation should lead to a better IR system.

Nevertheless, history charts many examples where careful evaluations led researchers to incorrect conclusions. For example, in early iterations of the TREC Web Track, several teams experimented with hyperlink data but concluded it was at best not helpful and more likely made things worse [49]. Hyperlink data is in fact extremely helpful in web search, but the tasks and relevance criteria used in TREC at the time were not representative of web usage. Similarly, a poor proxy measure led Netflix to optimise their recommender incorrectly [46, 101]. In the other direction, search systems clearly improved in the twenty years to 2009 despite claims that some metrics were not improving overall [6]. The problems were with *validity*, the extent to which an evaluation really reflects what we care about.

The emergence of retrieval-augmented generation (RAG) and question answering (QA) systems further expand what is being evaluated. No longer are documents merely retrieved; systems coordinate searches, synthesise information, and present answers. We might consider properties such as grounding, coherence, or accuracy. Outputs of RAG and QA systems are increasingly assessed by machines (e.g., LLM-as-a-judge [125]). Further, in the case of a multi-stage process, we might care about some intermediate quality even if that's not directly visible to the searcher. Such properties might include query generation, retrieval quality, synthesis accuracy, and presentation effectiveness.

We define an "evaluation" as being *an assessment of some quality* of an information access system. In general, evaluations can take many forms: predictions of relevance, utility, accuracy, hallucination, coherence, or anything else; an aggregation or a single thing; pointwise or pairwise or listwise assessments; with data taken from first-, second-, or third-party assessors and from a range of instrumentation [105]. Researchers and system designers build and use evaluations to understand a target quality, assuming that an improved evaluation (apparent effectiveness) predicts some real improvement in experience (actual effectiveness). The assumption holds *as long as the evaluation protocol is valid.*

Figure 1 illustrates this idea. We care about the *actual effectiveness* of a system, the degree to which it agrees with a specified target. An evaluation protocol gives us *apparent effectiveness*, a proxy for actual effectiveness. The *validity* of this evaluation is the degree to which it agrees with the target: the more valid, the closer our apparent effectiveness is to actual effectiveness and the better the proxy. Of course, we can choose multiple evaluations, and each

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808538

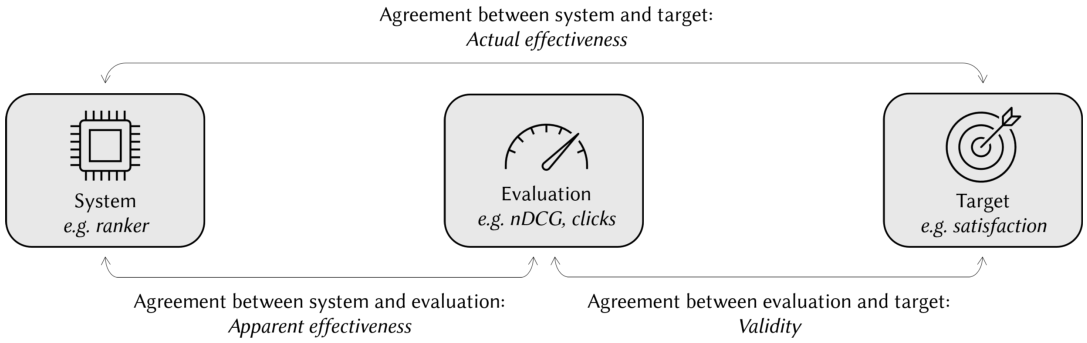

Figure 1: The relationship between system effectiveness (actual and apparent) and validity. *Apparent* effectiveness is a proxy for *actual* effectiveness. A *valid* evaluation means a better proxy.

will furnish some score (some apparent effectiveness) and will have some predictive power (some degree of validity), although the *actual* effectiveness will be the same in each case.

How large is each gap? The left-hand gap can be measured: apparent effectiveness tells us exactly how closely a system agrees with some evaluation. The right-hand gap is harder. Even in the case of simple measures of document relevance or ranking, we know it can be substantial (Section 9). In the case of more abstract targets such as utility or knowledge gain, and even more markedly in the case of targets for newer technologies such as RAG, we do not have a good idea what the targets even *are* (Section 5).

Substantial effort has been devoted to creating new evaluations and especially to improving apparent effectiveness—closing the left-hand gap in Figure 1. There has been relatively little effort in improving validity and closing the right-hand gap, but we argue that validity's importance is only growing: it helps us choose evaluations, helps us make progress as metrics saturate, clarifies the role of new evaluation techniques, and ensures we are optimising towards what's actually important.

We present a systematic framework for assessing whether evaluations actually measure what they aim to measure. The framework is organised around six core principles that any validity assessment should address, supported by design guidelines and rubrics that can be applied in any order.

Perspective. This paper is informed by our experiences evaluating tools ranging from rankers to conversational agents, in large industrial and academic settings including in initiatives such as TREC, in smaller experiments, and with a wide variety of evaluation styles. We will emphasise the right-hand side of Figure 1, which is to say *the validity of our evaluations*: how should we talk about validity, how can we determine that an evaluation or protocol is valid, and how can we improve validity over time? We start, however, by looking where many IR papers focus: the gap on the left-hand side.

2 The Gap Between System and Evaluation

Retrieval and evaluation are the same thing at heart: in both cases we ask "how good is this result in this situation?", albeit in different contexts. That is, the system that we are measuring and the process performing the measurement are trying to predict the same thing.

Soboroff [99] correctly observes that predictions from the evaluation need to be somehow better than those from the system—more valid, closer to the phenomena we really care about (utility or task completion, for example)—or measurement collapses. This difference between the quality of the system and of the evaluation process is critically important. In Figure 1, we want the evaluation to be closer to the target than the system is: i.e., we want the evaluation to be a better predictor of searcher experience than the system is.

Historically we have maintained this difference, even as retrieval has improved, by giving the evaluation much more resources. For example, we insist that retrieval runs in interactive time ($<$ 1s) while evaluation can take days; retrieval uses only keywords while evaluation has task descriptions; and most importantly retrieval has used machines and evaluation has used people. Those extra resources are reasonably assumed to lead to evaluations much closer to the searcher's experience. The task for system designers is to bring the system closer to the evaluation, even with this imbalance: that is, to produce better scores and to close the left-hand gap.

2.1 Shrinking the Left-Hand Gap

Many papers in IR assume the evaluation is correct, then operate on the left side of the figure to increase apparent effectiveness (e.g., metric scores). The standard practice is to choose some baseline, and an evaluation protocol, then introduce a new approach that outperforms the baseline; typically, we want to see statistically significant improvements over the baseline, on multiple data sets. Rarely, we report improvements on several unrelated evaluations. This approach has led to unequivocal improvement in IR systems, and is encouraged by leaderboards and publication bias.

There are problems with this approach: notably, selection bias means that published papers are often using weak baselines or are just lucky [6]. Nevertheless, we are slow to adopt new baselines and only big steps forward such as BM25 [87], or neural methods [69], have changed what is commonly used.

2.2 Saturating Scores

As systems continue to improve, the gap between systems and evaluation may well be smaller than before: scores will increase, and start to saturate. This could lead to trouble measuring. For the time being, we do not see this in practice since practical systems

are still limited compared to evaluation (in commercial systems, evaluation still takes hours or days while retrieval is measured in milliseconds), and arguments from efficiency suggest complete collapse is some way off, but some risks are apparent.

One particular risk is that as Large Language Model (LLM)-based assessment gets more efficient, it becomes more and more tempting to incorporate it holus bolus in running systems: that is, to use the same LLM-based machinery to produce a result and to evaluate it. A pre-LLM parallel would be to ask the assessor for a query to also come up with a ranking, then evaluate the ranking using that assessor's labels, and conclude that the assessor's own ranking was better than the baseline.

This makes it possible to "cheat" the evaluation. Even allowing for variance in LLM output this will reduce the left-hand gap, and we can drive it close to zero as we build arbitrarily expensive systems. This puts us at the limit of what we can evaluate, perhaps even stalling improvements while we develop a new way to evaluate [99].

This is not just a theoretical concern. Knowing that the LLM-based UMBRELA tool [109] was being used in the TREC 2024 RAG Track, Clarke and Dietz [24] submitted a system which re-ranked results according to UMBRELA's output. Unsurprisingly, since the "retrieval" system and evaluation used the same process, the system agreed with the evaluation and rated well. With less noise in the LLM, and more information about the evaluation protocol, it should be possible to achieve a perfect score. This has led several authors to urge caution in LLM-based assessment [24, 36, 99].

We observe two possible solutions, and one false "solution".

No LLMs. A false "solution" to the saturation problem is to outlaw LLM-based evaluation, only using human-based (or non-LLM-based) evaluation to measure apparent effectiveness. This is a false solution for two reasons: first, while addressing the problem with LLMs it does not tell us where we should draw the line (is a system with an expert human in the loop also out of bounds? If we have a medical professional providing assessments, and a similarly qualified medical professional helping sort documents, is this allowed? Are past TREC interactive runs useful?). More importantly, it is a non-solution because we are not interrogating the evaluation itself. Any evaluation, no matter how it is built, can be imperfect; as experimenters, we should investigate the role of LLMs, or any other new process, in the hope of richer and more powerful evaluations that can close the right-hand gap.

Cost and practicality. Another response is to consider the cost of the IR system, and how difficult it is to deploy. The IR field is known for building practical systems that influence, or are directly adopted by, full-scale production search services; for example, we often want to make ranking as cheap as possible so we can scale to millions or billions of documents. If arbitrarily increasing apparent effectiveness—increasing scores on standard metrics—needs an arbitrarily expensive system, it may not be viable. We may instead accept research that closes the left-hand gap as far as possible, but works within practical bounds. This may rule out adopting evaluation protocols inside systems, particularly as these protocols get more sophisticated.

Acceptance. A final response is to accept that systems will continue to improve on whatever metric we use, and may eventually saturate the metric—even closing the gap completely in some cases, making the system identical to the evaluation. As a thought experiment we can consider what it means if the system perfectly agrees with the evaluation, but the evaluation also perfectly agrees with our target. In this case, we have "solved IR": we have built a machine that completely agrees with our goal!

Alternatively, if the metric is saturated but does not agree with the target—that is, we score 100% but our goal is not perfectly satisfied—we can still improve the evaluation process to better agree with the target. Changing the evaluation will both open up the left-hand gap (i.e., scores will no longer be saturated), and eventually improve the system overall as we optimise towards a new and better metric. *Once apparent effectiveness is starting to saturate, we can improve actual effectiveness by improving validity.* In fact, this may be the only tool we have.

No matter how good our systems appear, evaluation problems are unlikely to go away soon: it is fairly easy to find questionable assessments, from humans, LLMs, or elsewhere, meaning there are still plenty of improvements we can make to our protocols. More importantly, it is still not always clear what we should be measuring in the first place.

3 The Gap Between Evaluation and Target

An evaluation, in Figure 1, is a proxy for a searcher or other expert—ideally, it will judge the system as a searcher would. Unfortunately, the intents and the judgements of searchers are rarely observable. Usually we need to look at signals such as prompts or queries to guess at the searcher's true intent; similarly, we need to look at interactions such as clicks or abandonment to guess at judgements. We can ask the searcher to more clearly articulate their thoughts, but this compromises usability, and in some cases the searcher may not be fully aware or able to articulate their true intent. We could instead ask other humans or LLMs to try to guess at the searcher's thoughts, but they may draw incorrect conclusions. This can lead to a systemic gap between the searcher's intent and the system we use to evaluate whether it was met.

This gap is acknowledged in the literature on IR metrics, although not often discussed. One positive example is Zobel [126], who introduces *karpos* as "the degree of success in achieving a qualitative aim", or in our context how well a system does whatever it is we want it to do. He also identifies a *lacuna* between what proxies indicate, and karpos; this lacuna is a close parallel to what we identify as the right-hand gap between what an evaluation tells us is good, and what we would say. Zobel goes on to claim this gap cannot be closed. This is of course a reasonable hypothesis, but leaves open important questions of how small a gap is "small enough" and what we can do to measure it. We take up these challenges here.

3.1 Third-Party Judges and "Gold"

Evaluations abstract and stand in for searchers when we test IR systems, meaning any errors or bias will lead to unreliable conclusions. In particular, if the evaluation deviates from what a user values—i.e., the evaluation is invalid—then a system users think is better might be evaluated poorly compared to a system users think is worse. Speaking in particular of TREC-style efforts, Soboroff [99] notes

that the "relevance judgments barrier is a fundamental limitation of evaluations that measure systems against ground truth".

We therefore need evaluations—whether based on TREC-style judgements, or on telemetry, or anything else—to represent the true quality of a result, to a properly qualified person in a real situation at a real time: we want them to be perfectly valid. Borrowing from Bailey et al. [10], we will define a "gold" assessor as someone with topical expertise and who had the information need: that is, a real user of the system (or a potential user, or an expert in the thing we are trying to measure) with a real need. Their assessments (which we will also call "gold") are as good as we can get, and are the standard against which we should measure.

In practice, there have been two ways to collect gold assessments. One is the TREC model, where assessors are contracted first for topic development, then label system outputs some time later [112]: this ensures the people providing assessments know exactly what was meant by each query, and are in the best position to judge systems. A more common model is to collect feedback from users of a running system: Kim et al. [59] and Thomas et al. [105] describe several feedback mechanisms for web search, for example, but we can also use surveys or apply other instruments in the lab.

Collecting gold assessments can be expensive. Soboroff [99] describes using six contractors for 2–4 weeks for a TREC track, which typically only has a few tens of judged queries. To evaluate over many documents, or information needs, or to make fine distinctions, we need large-scale data; and as always, we want to get as close as possible to gold assessments. The most common response is to use third-party assessors, i.e., assessors who do not have genuine information needs or sometimes even topical expertise [44, 64]. Bailey et al. call these "silver" and "bronze" assessors, respectively.

3.2 LLMs are Third-party Judges

We therefore have two choices: we can use a small amount of first-party "gold" assessments, and operate with real restrictions on scale; or we can use third-party "silver" or "bronze" assessments, and scale to much larger collections or much more diverse tasks but at some cost to validity. Demands for scale have generally led to the second option, including in influential cases such as many TREC collections.

If we choose not to rely exclusively on gold quality, we can use any process we like to generate assessments: crowd workers, researchers themselves, and graduate students have all been used and there is a good body of research on instructing and working with crowd assessors in particular [34, 65]. Increasingly, assessment is being done with LLMs, and especially by LLMs constructing labels for query-document pairs ("qrels" in TREC terminology) or RAG responses (i.e., LLM-as-a-judge [125]).

There is evidence that LLMs can produce qrels which agree with those produced by people, at vastly reduced cost and time. They have been used at scale in commercial systems [107] and have many people interested in practical possibilities [for example 26, 35, 45, 97]. LLMs are prone to documented biases and blind spots [e.g. 3, 97, 102] and are sensitive to the particular wording of a prompt [107, 122] and the order that examples are presented [122]. Further, it is not (yet) clear how good LLMs are at truly specialist evaluation: in our experience they detect spam well, for example, but they seem poor at evaluating medical text [103]. These or similar charges, however, can be levelled at any third-party judge—certainly crowd workers have biases and shortcuts [8, 57, 93, 96, 106]—so we should be able to discuss validity without pre-judging any particular implementation.[1]

3.3 Disagreement and Validity

Gold assessments give us the means to measure not just systems, but evaluations themselves. By comparing our assessments to gold, we can measure the error induced by the assessment protocol itself.

This is crucially different to merely showing that two sources (dis)agree. A common argument against using LLMs, for example, is that run-level evaluations based on LLMs do not look the same as those based on human labels (e.g., considering Kendall's τ over several systems), or evaluations have trouble distinguishing between the best-performing systems [38, 111], or they show unexpected disagreements [24]: that is, we get different conclusions depending whether our evaluations use third-party humans or LLMs.

If two evaluations come to different conclusions, however, we cannot decide that one is right and the other is wrong *unless we measure the validity of each*. These objections only hold if we can show that LLM-based evaluations—or evaluations from any new protocol—are actually *wrong* more often than are human-based alternatives. That is, we cannot just say one process is different and therefore bad; we need to show that one process is less *valid*, that it produces assessments counter to what we really believe. Again: *if we haven't tested the validity of an evaluation, we have no grounds for claiming one is "right" when another is "wrong"*. As Soboroff notes: "the assessor is not all-knowing, all-seeing, all-reading with perfect clarity. Assessors make mistakes" [99].

The best use for discussing (dis)agreement between evaluations is that if two evaluations agree, and we know something about the validity of one, we can draw conclusions about the validity of the other. As a special case of this, if one evaluation is from a gold source—a real searcher, in a real context—and another evaluation agrees, then the second is most likely valid. (If evaluation A and the searcher always agree, such as when A is "gold", then any time evaluation B agrees with A it also agrees with the searcher.) This is the argument from Thomas et al. [107] and the basis for their "superhuman" claims: the LLM agreed with gold labels more often than did their third-party (human) judges.

4 Talking About, and Testing, Validity

We have established that the validity of our evaluations is important, both for building better metrics and building better software. It helps us choose between evaluation protocols, it gives us confidence that we are actually improving the search experience, and it is a response to fears of model collapse. We will now turn to ways we can discuss, test, and measure the validity of our evaluations.

The term "validity" covers many distinct properties, all of which are desirable in our abstractions and measurements.[2] For example

[1] One risk that is unique to LLMs is contamination, or the possibility that the model has memorised gold sets or other relevant data during pretraining. This need not undermine validity—gold sets are ground truth after all—but it may point to later trouble generalising, and it may undermine validity for evaluation on adjacent tasks.
[2] Some of these properties, unfortunately, have been given different names by different authors at different times.

Table 1: Principles for improving evaluation validity.

(1) Define the target construct.
(2) Consider theoretical consistency. *Use expert review, factor analysis; test direction and relationships; test components.*
(3) Consider reliability. *Use A/A testing and resampling. Consider the effect of aggregation.*
(4) Consider concurrent validity. *Look at related evaluations.*
(5) Consider predictive validity. *Test against gold labels, carefully sampled; build controlled degradations.*
(6) Examine results carefully. *Accept, but examine, some residual invalidity.*

we could (and should) consider *external validity* (can we learn about new things, can we generalise to new circumstances?), and *concurrent validity* (does this evaluation agree with others that purport to measure the same thing?). The most important for our purposes is *construct validity*, which is pertinent whenever we are trying to measure something abstract like satisfaction or relevance.[3] This asks: are we measuring a real thing? Is it the thing we care about? Does this assessment provide an accurate measure of the construct we're interested in?

There are several core principles we should expect of any evaluation (Table 1). Failure to adequately address any of these will compromise validity. We discuss each of these principles below, in the context of information retrieval and RAG systems, along with techniques we can recommend in each case.

5 Defining the Target Construct

The hardest, but most important, part of assessing validity is establishing and defining the construct that we claim to measure. Target constructs are often outlined when evaluating new tasks [4, 110, 120], but the justification of their selection is less often provided. As we create entirely new styles of generative information access, we need to be sure that selected target constructs are valid.

Any target may be a unitary construct or a composite of multiple constructs; each could be latent or directly observable. A construct may further be a property of a system or its output (e.g., factual correctness of generated text, citation accuracy), a property of a user experience or design goal (e.g., whether users felt their question was answered), or a combination of both (in which case, the theoretical relationship between them must be specified). System properties and user outcomes may diverge: users can be satisfied with incorrect answers, for example, or dissatisfied with correct ones. A validity framework must account for what is being measured and when to expect any divergence.

An operational definition of the target construct should be such that an independent researcher could apply it consistently, but several common targets in IR research are notably ill-defined. "Relevance" covers a great many distinct concepts [13, 91, 92], not always carefully distinguished, while "utility" and "satisfaction" are entirely under-examined [18].[4] Ambiguity at this stage propagates through any later validity assessment. Without specifying what, exactly, we are trying to measure it is hard to trust that our evaluations are meaningful.

Despite the importance of establishing and defining target constructs, there are few published studies that actually do this, one example is detailed in Garcia-Gathright et al. [42]. The work was conducted after realising that a commercial recommender system was being used in ways that were not anticipated in the original system design, and a new set of target constructs had to be established. Garcia-Gathright et al. detail ways in which qualitative and quantitative researchers worked together to create new constructs, and to identify ways in which the constructs could be measured with online metrics. Seeing further examples of such work would be of value to our community.

Boundaries. In our experience, it is useful to consider the boundaries of the evaluation. For example, we can describe what a low-quality response looks like, and we can generate clear positive and negative examples as well as identifying ambiguous cases and specifying how they should be handled. In practice, we have found that articulating the target construct can be difficult, and it is often helpful to start with examples of things that are bad, as this can illustrate examples of things that are good.

6 Theoretical Consistency

A followup question could be: do the concepts in the evaluation—the inputs to the process, the entities it employs, the properties it measures—align with what we would expect? Does the evaluation agree with the relevant theory? *Theoretical consistency*[5] assesses whether the metric behaves in accordance with expectations about how quality should work. An evaluation that sometimes moves in the wrong direction relative to theory indicates fundamental validity problems, or that the theoretical understanding is flawed. Textbook methods include expert review of concepts and empirical testing of relationships between constructs [32, 100].

This idea is well-established in many fields. The Organisation for Economic Co-operation and Development and the European Commission, for example, insist that "the choice of indicators must be guided by the theoretical framework" [79]; Spielman et al. [100] ask for "clear correspondence between the index's conceptual framework and measurable inputs".

This seems almost a trivial step—why would we have an evaluation that violates our theories?—but has been relevant even to fundamental IR measures, for example with debate about whether they have a coherent user model [52] or whether utils are interval data [39, 75]. IR is a pragmatic field without a lot of generally-accepted theory: certainly not of some important latent variables like "relevance", which has been defined more than once but where definitions have not overly influenced practice [13, 91].

As IR practitioners, we tend toward pragmatic and ad-hoc definitions, or indeed no definition at all, but there have been some examples of this kind of enquiry. For example, Liu et al. [70] and

[3] The term "construct validity" is largely due to Cronbach and Meehl [32]. See Lissitz and Samuelsen [67] for a good discussion of past and current understandings of construct validity and related concepts.

[4] One counter-example is "engagement", carefully examined by O'Brien [77].

[5] Cronbach and Meehl [32] call this "construct validity", although that term has come to cover more ground.

Wanner et al. [114] ask whether factuality measures for generative models should merely measure precision, or whether a "good" response should also be complete (high recall) and whether different claims should carry different weight. Moffat [74] suggests axioms for ranking metrics, based on an assumed theory of relevance: for example, a ranking should not be worse if we merely add more results to the end. From the other direction, Moffat et al. [76], Carterette [17], and Jones et al. [55] all invite us to consider the user models implied by common metrics. These are a useful start but, in every case, need grounding in what a user model or a good response *should* look like.

As IR practitioners, we tend toward pragmatic and ad-hoc definitions, or indeed no definition at all. If we do not know what it really is we are trying to measure, it is hard to argue that we are measuring correctly.

Component contributions. Complex latent constructs, such as "satisfying a user's information need", might be measured with a composite that combines multiple component signals. These components should align in the expected direction. This is not always the case, as can be shown by examining loadings in a factor analysis or by checking the rules of combination [100]. If components work against theoretical expectations, the aggregation method may be masking fundamental validity problems.

7 Reliability

Reliability assesses whether the evaluation produces stable, reproducible results. If we see substantially different assessments under minor methodological variations, the evaluation is unreliable. This is not the same thing as validity (a measure can be reliable but invalid), but does serve to limit validity (an unreliable evaluation cannot be valid). Any evaluation involving people or LLMs is nondeterministic to a degree [94], which is amplified by chains of models or agents, so reliability is a critical factor.

If different implementations of the same metric on the same data yield strikingly different results, the metric lacks the reliability required for valid measurement. Similarly, repeated random samples of the same system or multiple systems should not change the perception of quality or ordering of the systems.

Methods to determine reliability can include A/A testing, which measures test-retest reliability by simply performing the same evaluation over the same data twice [61], and resampling, which measures variance by resampling from some suitable population and seeing the extent to which the evaluation changes [e.g., 15, 43, 86]. This simply means we need to repeat across samples, or re-run the same evaluation, and note the variation from run to run. If there is wide variation, this is a possible source of invalidity: things might look different on a slightly different sample, and an evaluation which is unreliable leads to decisions based on luck rather than evidence.

An alternative approach is making small or inconsequential changes, to illustrate the robustness of the evaluation to changes in the mechanics. LLM judges, for example, can be sensitive to small changes in the text of the prompt [84, 107]. Chatterjee et al. [20] even show different responses from single-character changes in prompts, which exposes the uncomfortable possibility that "fixing" a prompt could make it worse. "Sweeping" across prompts, examples, guidelines, models, or other components is likely to find more- or less-reliable alternatives. Interest in LLM variance has led to methods such as POSIX [20], TARa@N [7], and FormatSpread [95], but we can also subject our evaluations to standard tests such as ICC [62], generalisability theory, or even simple ANOVA. We look forward to seeing more of this work.

Scale sensitivity. If the evaluation aggregates across units (questions, documents, users), we must assess whether changing the aggregation level changes conclusions. For example, responses that individually appear high quality might look different aggregated across a topic, a session, or a conversation; rankings might not be stable across different sample sizes. At Microsoft Bing, we found A/B tests gave results that differed by 20% if they were aggregated by interaction, by session, or by person, suggesting the metric or (more likely) the sampling process was unreliable.

These kinds of grouping and ordering effects are well known (see, e.g., Simpson's paradox). However, they raise questions about the validity of an evaluation to the target. These kinds of changes can occur because the concept that we are trying to measure is scale dependent, and changing scales alters the definition of the concept. Alternatively, it could be that the concept is scale-specific and changing scales does not actually make sense: something that is valid at, say, the query level might not be valid when aggregated to the conversation level.

8 Agreement with Other Indicators: Concurrent Validity

Concurrent validity [32, 67], *criterion validity* [30] or *external consistency* [100] is the extent to which two instruments, purporting to measure the same things, do in fact align. *Convergent validity* is similar, when two instruments measure similar things but where we might expect some differences. For our purposes, the two concepts are similar. We commonly have two measurements, or two ways of taking a measurement, claiming to measure the same construct, and they should correlate. For example, we may have a new metric that we believe predicts searcher satisfaction: it would be reasonable to ask whether it correlates with the Net Promoter Score.[6] For a new metric which we think overlaps with, but is different to, topical relevance, we might ask about correlation with TREC-style labels (and expect high, but not perfect, agreement). Zhang et al. [124], Azzopardi et al. [9], and Lipani et al. [66] use a similar idea to argue that some metrics (e.g., RBP, IFT, sRBP) have better models of searcher attention (correspond better to other indicators) and are therefore more informative than alternatives; Turpin and Scholer [108], to the contrary, demonstrate that MAP does *not* correlate with time on task nor the number of relevant documents searchers can find, i.e., has low convergent validity. Other research has suggested both that offline metrics may be valid [e.g., 90], valid but insensitive [e.g., 2], or simply invalid [e.g., 43, 51, 68], depending on the other indicator(s) chosen and the research setting.

Smith [98] points out problems with this approach. First, if we expect our new evaluation (call it A) to correlate with some other indicator (B) and it does not, we cannot conclude that A is invalid. It

[6] https://fortune.com/longform/net-promoter-score-fortune-500-customer-satisfaction-metric/

is of course possible that B is invalid instead, or that the supposition of correlation is wrong. Conversely, if A and B *do* correlate, we cannot conclude that A is valid—for example we could be measuring something else entirely, C, which just happens to correlate with B. As a practical illustration, we might ask whether a relevance metric agrees with observed behaviours such as clicks. This is a weak signal (clicks are sparse), but also potentially misleading: a metric that agrees with clicks might not correlate with relevance, but both the new metric and the clicks might correlate with something else, such as having key terms in the title [119]. In this example, both the new metric and the clicks would in fact be *in*valid in the same way—A and B correlating with each other, but not with relevance.

Even if evaluations agree, then, the critical question remains: do they correlate with user outcomes or other targets we care about? This motivates *predictive* validation.

9 Agreement with Target: Predictive Validity

Predictive or *empirical* validity assesses whether the metric predicts observable outcomes it should theoretically predict. A metric with strong theoretical grounding and internal consistency, but no relationship to user outcomes, has questionable practical value.

To test empirical validity, we need to identify outcomes the metric should predict based on its target construct, then test whether it does. A simple approach is to measure the correlation between an assessment and some relevant gold. This is exemplified by Chen et al. [21], who used offline metrics in the CWLA family and computed correlation with self-reported final satisfaction. The degree of correlation was interpreted as an indication of the quality of each metric. This is a loose indication, since CWLA metrics report information gain rather than satisfaction (this is a test of *convergent* validity), but nonetheless Chen et al. reported moderate correlation.

The clear advantage of using gold assessments is that, if collected correctly, gold assessments are correct. (They may not measure what we think, and they may not be useful; but by construction, they should be correct.) This approach has accordingly been widely used in the literature, with broadly good correlations between standard offline metrics and explicit assessments from lab studies [22, 23, 53, 117, 121] and has been used to justify new metrics [e.g., 123].

9.1 Constructing Gold Assessments

Sometimes our policy goal is such that we can construct results of known quality, without involving "real" participants. This is practical only when there is a well-defined objective such as "newer is better" (and where this is supported by data such as document age). Concepts such as relevance, or satisfaction, are subjective and constructed examples are not true gold—it is still possible for a valid assessment to disagree in these cases. This means the technique is useful for only a subset of concepts we care about. In our experience, it is still useful to construct assessments of more subjective concepts such as relevance, but since these are only silver (or even bronze) judgements we must be careful when interpreting agreement. In any case, constructing assessments has the advantage that we can generate any number of examples without involving searchers, even with entirely synthetic data, while still being confident we are correct according to our policy.

9.2 Agreement with Pairs

Instead of assigning concrete (pointwise) values, it is often easier to construct pairs of results where we know one is better than the other. For example, for news search we might always want to return the latest version of a story; we know by construction that a newer story is better than an older one on the same topic, even if we do not know exactly how good either story is by itself. An evaluation is valid to the extent that it reflects this difference. This is the "known groups" process, introduced by Cronbach and Meehl [32] (who call it "group differences") and discussed by Hattie and Cooksey [48].

As a special case of this we can employ *controlled degradations*. The idea is the same: if we can construct two outputs, one known to be better than the other, this gives us a test of our evaluation protocols. Since we cannot make a system arbitrarily better,[7] we can instead make one worse in a controlled way. Examples will depend on the design policies but could include injecting known off-topic documents, adding factual errors; removing known-good parts of the result; introducing artificial delays [5, 14]; and so on. In web search for example we have degraded features such as spell correction, synonym handling, and click data to test relevance labelling. In conversational RAG, amongst other things, we have manipulated the style of generated text and removed sources to test measures of output quality and grounding.

Dmitriev and Wu [37] used a similar idea with an "'interesting' corpus" of experiments at Bing. These were "representative experiments from different feature areas, 'learning' experiments that were run for the sole purpose of understanding user behaviour, experiments that had known bugs negatively impacting users, etc". Each experiment in the corpus was reviewed by both the original experimenter and by evaluation experts, looking at a suite of data including user studies and first-person feedback, to decide whether the experiment was good or bad for searchers overall. These high-quality labels could then be used in the same way as degradations: an evaluation should be able to tell "good" from "bad".

Unlike sampling gold assessments from live use, degradations let us control both the nature and scale of a difference: we can ensure that a pair of results differs in some particular aspect, and often we can control how big the difference is (for example, we can control the extra latency or the amount of spam). This means we can test not only whether the evaluation chooses correctly, but also how sensitive it is. If the assessment scale allows, we can also test the nature of its response—for example, whether the measured difference scales linearly with the real difference or whether there is some threshold. The combination of validity and sensitivity tests, on controlled features of the search result, make controlled degradations particularly useful. Since degradations use the running system, they also provide a partial solution to "stale" gold labels.

9.3 Meta-Evaluation

Comparison with gold also lets us build *meta-evaluations*, giving measures for evaluation themselves [105].[8] Given a meta-evaluation, we can use it to refine and improve our evaluations in the same way as we use metrics like nDCG to improve systems [e.g., 69, 72]. To

[7] If you know how—please get in touch.
[8] Thomas et al. referred to "metrics" where we use "evaluation".

our knowledge the choice of statistic has not been explored in the literature, despite the common observation that different statistics give different results. We have several choices.

When we have pointwise evaluations, we have several good options. Besides simple accuracy [80], which requires a balanced test set, probably the most common statistic in this case is Cohen's or Fleiss's κ [28, 40], widely used to compare assessments [19, 31, 33, 38, 50, 82, 107]. Alternatives for categorical data include weighted κ [29] or macro-average F1; mean absolute error for interval data; r^2; or Krippendorf's α [63]. Of these, α has advantages in that it is robust to missing data and unbalanced gold, and is defined over categorical, ordinal, and interval scales.

Given pairs of gold assessments (Section 9.2), we want an evaluation to correctly distinguish "good" from "bad" and, where appropriate, estimate the size of the difference. Pairs are by nature balanced, so a simple approach which we have used to some success it to measure the area under the ROC curve (ROC AUC). This is easy to interpret, but for completeness it needs pairs spanning the whole range of possible values. It has also been useful to report the average difference between the scores for the better and worse result. The larger the gap, the more sensitive the evaluation.

It is possible for individual evaluations (e.g., scores for responses) to generally agree with gold data, but listwise assessments (e.g., a ranking of responses) to disagree: this can happen, for example, if individual errors are concentrated in such a way as to influence one system's score more than others' [71, 111][9]. There are many rank correlation measures which are useful to test for this. Kendall's τ is most commonly used, but has two disadvantages: it is insensitive to the scale of disagreement, and it treats mis-orderings the same whether they are at the head of the list (the most effective systems, for example) or the tail (the least effective). In practice we may be more interested in larger errors and are likely interested in only one end of the ranking. Better measures include τ_{AP} [118], rank-biased overlap (RBO) [115], Pearson-rank (ρ_r) [41], and compatibility [25], which variously are sensitive to scale (ρ_r), are head-weighted (τ_{AP}, RBO, ρ_r, compatibility), or take one ranking as a "correct" reference (τ_{AP}, ρ_r, compatibility). All have seen some use in the literature.

Just as we should choose an evaluation to reflect what's important to the system, we can and should choose a meta-evaluation to reflect what's important in our evaluations. For example, if the evaluations are used to compare the current version of a system to some candidate replacement, it is important that we can distinguish small gaps between systems; a pairwise measure might be best. If we are selecting amongst many alternatives, a top-weighted list measure might be appropriate. If our assessments are used to debug search failures, it is important that every individual case is correct and an individual-level agreement measure might be more useful. Having more than one meta-evaluation gives more insight than any one alone. In practice at Microsoft, we have found that pairwise meta-measures give us an evaluation that supports experimentation, and result-level accuracy gives us an evaluation that better supports debugging, so we use both.

[9]There were earlier concerns this may have happened when neural systems used TREC data: the neural methods may retrieve previously unjudged documents, according to the argument, and since these would be considered non-relevant the evaluation would be biased against new systems. This concern was premature [56, 113], but was certainly reasonable.

9.4 Sampling and Controlling for Confounds

Of course, gold labels will only represent a small fraction of all possible use cases: we cannot produce gold labels for all people, all information needs, all interaction modes, at all times (or there would be no retrieval left to do). Equally obviously, if our gold labels do not represent some subset of uses then we cannot be sure our evaluations will, either. Any empirical validation must control for confounds from (at least) people, tasks, information, and responses.

For example, different groups of people will tend to have different opinions, especially on subjective tasks, and these differences do translate to machine-learned models [1, 81]. We also see differences according to demographics [58] and expertise: expert users may evaluate quality differently than novices [60, 73, 104, 116].

In conventional IR systems, there is also great variance due to topic (the information need) [86] and query wording (the way the need is expressed). Rashidi et al. [83] demonstrate that TREC's standard queries give different system-level evaluations than do queries from crowd workers, even though the queries are on the same topics. Bailey et al. [11] similarly saw more variation in evaluation scores due to query wording than due to topic or system. Some questions, of course, are inherently more difficult to answer well at all, and some topics have better source material available.

We also note that the particular responses we are evaluating, or even the sorts of objects we are evaluating, may change over time. In the case of ranked retrieval, we take a query and return a list of documents, and shared testbeds like TREC are useful long-term precisely because this format has changed little over decades. With RAG systems, that is less true and both the inputs and outputs are changing rapidly (inputs can be longer as models have larger contexts, can now be grounded in richer data, and can implicitly reference remembered facts or past interactions; outputs are longer, more multi-modal, and denser with links). This can mean that our sample of gold labels may be "stale", no longer matching what we are actually evaluating, and are less useful for testing validity.

The lesson is that while we act as if evaluations based on gold labels represent a system in the abstract, across all people, needs, and times, in fact they represent only a subset or sample of possible uses (and of outputs too, in some cases, e.g., the stochastic outputs of LLMs). If that sample does not represent some real usage, the evaluations are not valid; further, we cannot even test their validity.

10 Examining Results: A Caution

The techniques above, especially when used in combination, can give us confidence that our evaluations are valid—that is, that they are measuring what we actually care about—as well as insight into improvements. In our experience, however, there is no replacement for informed people, knowledgeable in both evaluation and the search domain, carefully examining both "correct" cases and discrepancies. Most of the time, discrepancies can be attributed to noise or error, either in the evaluation or the meta-evaluation, but there can be trends which inform the next iteration of design. From time to time, the "gold" is in fact in the wrong, which can suggest

improvements in meta-evaluation and understanding system use or system policy. We can only distinguish these cases by looking.

A concrete example of this is in work by Alaofi et al. [3], who were investigating using LLMs to evaluate over query variations. The overall meta-evaluation looked reasonable: different language models had different characteristics, but performance roughly correlated with assumed model capability and most results were reasonable. However, close inspection of the differences showed that LLMs were being fooled by simple keyword matching, leading to discoveries about the current limitations of LLMs. These discoveries hinged on the original expert cross-checks.

Zobel [126] gives further examples where blindly accepting aggregates would miss important details: "[a]re there subpopulations of queries that seem to vary together? Are the results influenced by the chosen effectiveness measure? Is there a relationship to other artefacts, such as the number of known relevant documents? Are poor scores due to unjudged documents…?". Zobel identifies these as problems of ecological validity.

11 Residual Invalidity

Regardless of our effort, it seems inevitable that there will be some gap on the right-hand side of Figure 1: there will always be some disagreement between even the most sophisticated evaluation and the people who use our systems.

This is true even if we evaluate exclusively with gold labels: even if we ask the real searcher, at the moment they are searching, they may be unreliable. Scholer et al. [94] report that judges, given two near-identical documents and the same information need, agree with themselves only 76–85% of the time; Thomas et al. [106] report differences due to holidays; Shokouhi et al. [96] and Scholer et al. [93] report evidence for an anchoring effect; and so on. (See Azzopardi [8] for discussion of other cognitive biases we might expect to see.)

Further, there is often limited scope to improve validity since even gold labels often provide a limited signal. "Thumbs up"-style feedback does not include any explanation, for example, and questionnaires only answer the questions we think to ask. Without knowing *why* a label was given, it is much harder to close the gap.

We might also choose an evaluation for pragmatic reasons such as cost or time. An evaluation which is closer to a person might also be less explainable, or harder to debug and act upon [107].

Over-reliance on an invalid measure—even a "gold" measure—was at the root of problems with OpenAI's GPT-4o [78]. Shortly after launch, users reported the model was prone to sycophancy: "validating doubts, fueling anger, urging impulsive actions, or reinforcing negative emotions". It transpired that OpenAI trained and evaluated models on several measures, including expert feedback, safety evaluations, and thumbs-up/thumbs-down signals from pilot users. Thumbs-up signals are gold—they are from real users, with real uses—but not valid measures of OpenAI's targets. Focussing on these signals at the expense of others led to a model that failed important safety criteria. In response, OpenAI committed to "weighing both quantitative and qualitative signals".

Similar observations were earlier made at Netflix, where optimising toward clicks did not lead to the real target of subscriber retention [46, 101], and Facebook, where engagement data led to radicalisation [85, 88]. In each case, differences between a gold signal and what really mattered led to substantially poorer products.

It seems that some gap between evaluations and the real searcher is unavoidable. Of course, we still want to use evaluations as a tool to improve our systems. An obvious response is to use multiple evaluations, as different as is feasible, to measure the same construct—for example to use both TREC-style labels and clickthrough data to estimate the utility of results, or signals from both language and copy-paste behaviour to estimate the success of a conversational session. If this portfolio of evaluations are all somewhat invalid, but in different ways; and if we improve on most or all of them; then there is a much better chance that the improvement is real. This argument is similar to that for hold-out test sets in machine learning, for example: we do not want to over-fit to a flawed goal.

12 Pragmatics

As well as the explicit tests above, we can suggest design practices that help build and maintain validity.

Documentation. A validity assessment should include explicit documentation of the evaluation's valid operating range and known limitations. For the former, it is useful to specify the conditions under which the evaluation has been validated: relevant dimensions might include query or task types, response formats, user populations, topical domains, and scale. If the evaluation is used to compare systems (rather than individual responses), we should also document what assumptions are required for valid aggregation: can we average scores? what sample size is needed? should we weight different query types or user segments?

For the latter, we should identify conditions under which the metric breaks down or produces misleading results, especially if results mean something different in certain contexts.

Competing requirements. The validity of an evaluation must be weighed against the practical costs of construction maintenance. A theoretically ideal protocol that is prohibitively expensive to run may be less useful than a simpler protocol with acceptable validity. Cost here can mean time and compute (particularly for LLM-based judges), creating gold-standard reference answers or evaluation rubrics, scalability constraints, or infrastructure dependencies and their reliability. Evaluations that require human involvement are also constrained by ongoing training and calibration, budget, and expertise requirements and availability of qualified evaluators. Thomas et al. [105] discuss these and other tradeoffs.

Again, we should document the relationship between cost and validity. Can a cheaper approximation achieve acceptable validity? At what cost threshold does the evaluation become impractical for its intended use? Understanding these tradeoffs enables informed decisions about metric selection.

Transparency. Finally, a evaluation's validity is constrained by how well it can be understood and scrutinised. Opaque metrics resist external validation and make it difficult to diagnose when and why they fail. Can stakeholders understand what a given score means in practical terms? Are score differences meaningful? Can we provide explanations or rationales for scores?

We should also consider whether the evaluation can be independently replicated (which means full documentation of the methods

and accessible components), and the extent to which its behaviour can be audited and challenged: Can specific scoring decisions be traced back to their causes? Is there a mechanism for identifying and investigating anomalous scores? Can the metric be stress-tested by external parties to reveal failure modes?

Evaluations that work as black boxes—particularly those based on proprietary LLMs or undisclosed methods—present concerns because their behaviour cannot be fully examined or understood.

13 Discussion

In IR, we *optimise systems under some evaluation* and simultaneously try to *ensure the evaluation models real targets*. To build an effective system, then, we need to close gaps on two sides: the system should agree with the evaluation on what makes a good result, and the evaluation should agree with the target. If we can perfectly close these two gaps, we have "solved" IR, since the evaluation perfectly represents the target (optimal metric validity) and the system is perfectly optimised for the metric (optimal measured performance). This is true regardless of the nature of the system or of the evaluation protocol.

We suggest that IR, as a discipline, would be well served by closer attention to metric validity. Most obviously, this gives us confidence that we are optimising for the right thing. Second, this gives us a way to talk about evaluations and to help choose between them when alternatives exist.

Considering validity can also make for novel evaluation. For example, a detailed examination of searcher behaviour led to the realisation that the way people search is highly individualistic; even if they are searching for the same information, the way that searchers word their queries is different. Such individuality was missing from existing test collections, hurting validity. This lead to the development of the UQV100 collection, an entirely new type of testing resource [12].

Less obviously but most importantly, it forces us to think about what the "right thing" actually is. Asking about validity makes us consider what actually serves our goals, and what a real system could and should look like. We might agree that a good system is *not* simply one that maximises nDCG or some other standard metric (unless of course we have a narrow view of success), but to discuss validity we need to have a concrete alternative.

Recommendations. This paper argues for discussing, measuring, and understanding the validity of IR evaluation so we can make informed choices. This means at least:

(1) Being aware of the gap, i.e., being aware of challenges to the validity of our evaluations;
(2) Being explicit about what construct we think we are evaluating, and what theory (if any) underpins this;
(3) When proposing a new evaluation regime—a new metric, a new test set, or a new protocol for example—considering whether it is supported by theory, is internally consistent, exhibits convergent validity, and has predictive validity;
(4) Continuing to challenge evaluations with gold data sets, controlled degradations, and other techniques so we can understand the limits of our methods; and
(5) Continuing to examine cases individually, and thoughtfully, instead of relying on aggregates.

Continuing research. From the discussion above we can make some observations and suggestions for continuing research.

First, nothing here argues against the stereotypical IR study or paper, introducing a new system and showing improved numbers. Of course it is acceptable to show that given some evaluation (of reasonable validity) we can improve system performance.

IR is known for building deployable systems, where developers can pick up a textbook and learn how to build a practical system. Sometimes it is impossible to deploy a practical system that contains an evaluation, for example if it involves human judges or involves LLM assessments that are too expensive to deploy at scale. In other cases, the evaluation may have more information than the retrieval system: for example, it may use a detailed description of user intent. This leaves plenty of scope for traditional effectiveness papers, and we would suggest maintaining the IR practice of running multiple metrics and datasets in each paper, to prove that there was no overfitting to one metric. These principles hold true whether the paper uses a diverse set of LLM-based user simulation approaches, a number of human evaluations, or a mix of both.

Second, nor does anything here compel us to embrace new evaluation. It is acceptable to argue, for example, that LLMs-as-judges should not be used because of a validity gap; for example, by showing that crowd judges are better at modelling the concerns of real users than an LLM judge. One possible future is that faithfully modelling a searcher is impossible, and it is always preferable to use human assessments. Another possibility is that we recognise that the inherently interactive nature of IR is important, and it is better to capture this interaction using perhaps a hybrid of human (setting the goal and topic description) and artificial (simulating the user's preferences based on that goal/topic) methods. Even more likely is that there are multiple evaluations, and both of the above are valid. These tradeoffs, the "future of metrics" best practices, are determined in evaluation validity work, in the right-hand side of Figure 1.

Acknowledgments

We are indebted to many colleagues for enlightening conversations on metric design; we particularly thank Bhaskar Mitra, Doug Oard, Gianluca Demartini, Ian Soboroff, Marwah Alaofi, and Leif Azzopardi. We thank the anonymous reviewers for their useful comments.

References

[1] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proc. Workshop on Online Abuse and Harms*. 184–190.
[2] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. 2008. The good and the bad system: Does the test collection predict users' effectiveness?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 59–66.
[3] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be fooled into labelling a document as relevant (best café near me; this paper is perfectly relevant). In *Proceedings of the International ACM SIGIR Conference on Information Retrieval in the Asia Pacific*.
[4] Mohammad Aliannejadi, Zahra Abbasiantaeb, Simon Lupart, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2024: The Interactive Knowledge Assistance Track overview. In *Proceedings of the Text REtrieval Conference*.
[5] Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. 2014. Impact of response latency on user behavior in web search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 103–112.

[6] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 601–610.

[7] Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" LLM settings. arXiv:2408.04667 [cs.CL] https://arxiv.org/abs/2408.04667

[8] Leif Azzopardi. 2021. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the Conference on Human Information Interaction and Retrieval*.

[9] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: An information foraging based measure. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[10] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: Are judges exchangeable and does it matter?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 667–674.

[11] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 625–634.

[12] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A test collection with query variability. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 725–728.

[13] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the Association for Information Science and Technology* 54, 10 (2003), 913–925.

[14] Jake D. Brutlag, Hilary Hutchinson, and Maria Stone. 2008. User preference and search engine latency. In *Proceedings of the ASA Quality and Productivity Research Conference*.

[15] Chris Buckley and Ellen M Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 33–40.

[16] Chris Buckley and Ellen M. Voorhees. 2005. Retrieval system evaluation. In *TREC: Experiment and evaluation in information retrieval*, Ellen M. Voorhees and Donna K. Harman (Eds.). MIT Press, 53–78.

[17] Ben Carterette. 2011. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 903–912.

[18] Donald O. Case. 2012. *Looking for information: A survey of research on information seeking, needs, and behavior* (3 ed.). Emerald, Chapter The concept of information.

[19] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. 2006. A reference collection for web spam. *ACM SIGIR Forum* 40, 2 (Dec. 2006), 11–24.

[20] Anwoy Chatterjee, H. S. V. N. S. Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. arXiv:2410.02185 [cs.CL] https://arxiv.org/abs/2410.02185

[21] Nuo Chen, Jiquin Liu, and Tetsuya Sakai. 2023. A reference-dependent model for web search evaluation: Understanding and measuring the experience of boundedly rational users. In *Proceedings of the International Conference on World Wide Web*.

[22] Nuo Chen, Fan Zhang, and Tetsuya Sakai. 2022. Constructing better evaluation metrics by incorporating the anchoring effect into the user model. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2709–2714.

[23] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[24] Charles Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. In *Proceedings of the International Workshop on Evaluating Information Access*.

[25] Charles L.A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline evaluation without gain. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 185–192.

[26] Charles L. A. Clarke, Gianluca Demartini, Laura Dietz, Guglielmo Faggioli, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Ian Soboroff, Benno Stein, and Henning Wachsmuth. 2023. HMC: A spectrum of human–machine-collaborative relevance judgment frameworks. In *Frontiers of Information Access Experimentation for Research and Education*, Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr (Eds.). Vol. 13. Leibniz-Zentrum für Informatik. Issue 1.

[27] Cyril W. Cleverdon. 1967. The Cranfield tests on index language devices. *Aslib Proceedings* 19, 6 (June 1967), 173–194.

[28] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[29] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (1968).

[30] Frederick L. Coolidge and Daniel L. Segal. 2010. Validity. In *The Corsini Encyclopedia of Psychology* (4 ed.), Irving B. Weiner and W. Edward Craighead (Eds.). Vol. 4. Wiley, Hoboken, NJ, USA.

[31] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 282–289.

[32] Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52 (1955), 281–302.

[33] Tadele T. Damessie, Taho P. Nghiem, Falk Scholer, and J. Shane Culpepper. 2017. Gauging the quality of relevance assessments using inter-rater agreement. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[34] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *Comput. Surveys* 51, 1 (2018), 1–40.

[35] Laura Dietz. 2024. A workbench for autograding retrieve/generate systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1963–1972.

[36] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and guidelines for the use of LLM judges. In *Proceedings of the International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*. 218–229.

[37] Pavel Dmitriev and Xian Wu. 2016. Measuring metrics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 429–437.

[38] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*.

[39] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. *IEEE Access* 9 (2021), 136182–136216.

[40] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[41] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. 2016. Pearson rank: A head-weighted gap-sensitive score-based correlation coefficient. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 941–944.

[42] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 55–64.

[43] Gebrekirstos G. Gebremeskel and Arjen P. de Vries. 2016. Recommender systems evaluations: Offline, online, time and A/A test. In *Working Notes of CLEF—Conference and Labs of the Evaluation forum*.

[44] Lukas Gienapp, Tim Hagen, Maik Fröbe, Matthias Hagen, Benno Stein, Martin Potthast, and Harrisen Scells. 2025. The viability of crowdsourcing for RAG evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 159–169.

[45] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056. arXiv:2303.15056 [cs.CL]

[46] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems* 6, 4 (Dec. 2015).

[47] Kai Halttunen and Kalervo Järvelin. 2005. Assessing learning outcomes in two information retrieval learning environments. *Information Processing and Management* 41, 4 (2005), 949–972.

[48] John Hattie and Ray W Cooksey. 1984. Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement* 8, 3 (1984), 295–305.

[49] David Hawking and Nick Craswell. 2005. The Very Large Collection and Web Tracks. In *TREC: Experiment and Evaluation in Information Retrieval*, Ellen M. Voorhees and Donna K. Harman (Eds.). MIT Press.

[50] William Hersh, Chris Buckley, TJ Leone, and David Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 192–201.

[51] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. 2000. Do batch and user evaluations give the same results?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 17–24.

[52] Kalervo Jarvelin and Eero Sormunen. 2024. A blueprint of IR evaluation integrating task and user characteristics. *ACM Transactions on Information Systems*

42, 6 (2024), 1–38.
[53] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 285–288.
[54] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 133–142.
[55] Timothy Jones, Paul Thomas, Falk Scholer, and Mark Sanderson. 2015. Features of disagreement between retrieval effectiveness measures. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 847–850.
[56] Jaap Kamps and David Rau. 2021. University of Amsterdam at TREC 2021: Deep learning track. In *Proceedings of the Text REtrieval Conference*.
[57] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and S.M.M Tahaghoghi. 2012. An analysis of systematic judging errors in information retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 105–114.
[58] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.
[59] Jin Young Kim, Jaime Teevan, and Nick Craswell. 2016. Explicit in situ user feedback for web search results. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 829–832.
[60] Julia Kiseleva, Alejandro Montes Garc/'ia, Jaap Kamps, and Nikita Spirin. 2015. The impact of technical domain expertise on search behavior and task outcome. arXiv:1512.07051v1 [cs.IR]
[61] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18 (2009), 140–181.
[62] Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163.
[63] Klaus Krippendorff. 2022. *Content analysis: An introduction to its methodology* (fourth ed.). SAGE Publications, Inc.
[64] Matt Lease and Emine Yilmaz. 2013. Crowdsourcing for information retrieval: Introduction to the special issue. *Information Retrieval Journal* 16 (2013), 91–100.
[65] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J. Franklin. 2016. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge & Data Engineering* 28, 9 (Sept. 2016).
[66] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a user model for query sessions to session Rank Biased Precision (sRBP). In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 109–116.
[67] Robert W. Lissitz and Karen Samuelsen. 2007. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher* 36, 8 (2007), 437–448.
[68] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2122–2132.
[69] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.
[70] Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. VeriFact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. arXiv:2505.09701 [cs.CL] https://arxiv.org/abs/2505.09701
[71] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv:2303.16634 [cs.CL]
[72] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning LLaMA for multi-stage text retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2425.
[73] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How does domain expertise affect users' search interaction and outcome in exploratory search? *ACM Transactions on Information Systems* 36, 4, Article 42 (July 2018).
[74] Alistair Moffat. 2013. Seven numeric properties of effectiveness metrics. In *Proceedings of the Asia Information Retrieval Societies*.
[75] Alistair Moffat. 2022. Batch evaluation metrics in information retrieval: Measures, scales, and meaning. *IEEE Access* 10 (2022), 105564–105577.
[76] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems* 35, 3 (2017), 1–38.
[77] Heather O'Brien. 2016. *Theoretical perspectives on user engagement*. Springer, Cham, 1–26.
[78] OpenAI. 2025. Expanding on what we missed with sycophancy. Retrieved 21 November 2025 from https://openai.com/index/expanding-on-sycophancy/.
[79] Organisation for Economic Co-Operation and Development and Joint Research Centre of the European Commission. 2008. Handbook on constructing composite indicators: Methodology and user guide. Retrieved 16 December 2025 from https://www.oecd.org/content/dam/oecd/en/publications/reports/2008/08/handbook-on-constructing-composite-indicators-methodology-and-user-guide_g1gh9301/9789264043466-en.pdf.
[80] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. arXiv:2306.00176v1 [cs.CL]
[81] Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. 2022. How does the crowd impact the model? A tool for raising awareness of social bias in crowdsourced training data. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.
[82] Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds: Exploring model label variation. In *ProcWorkshop on Perspectivist Approaches to NLP (NLPerspectives)*. 19–30.
[83] Lida Rashidi, Justin Zobel, and Alistair Moffat. 2024. Query variability and experimental consistency: A concerning case study. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 35–41.
[84] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity in large language models. In *Proceedings of the European Conference on Information Retrieval*.
[85] Manoel Horta Ribeiro, Robert West, Raphael Ottoni, Virgílio A. F. Almeida, and Wagner Meira Jr. 2021. Auditing radicalization pathways on YouTube. arXiv:1908.08313 [cs.CY]
[86] Stephen E. Robertson and Evangelos Kanoulas. 2012. On per-topic variance in IR evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 891–900.
[87] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Text REtrieval Conference*.
[88] Kevin Roose. 2020. Rabbit Hole. The New York Times: https://www.nytimes.com/column/rabbit-hole.
[89] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.
[90] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 555–562.
[91] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 1915–1933.
[92] Tefko Saracevic. 2012. Research on relevance in information science: A historical perspective. In *Proceedings of the ASIS&T Pre-conference on the History of ASIS&T and Information Science and Technology*. 49–60.
[93] Falk Scholer, Diane Kelly, Wan Ching Wu, Hanseul S. Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 623–632.
[94] Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1063–1072.
[95] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*. 25055–25083.
[96] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and adjustment in relevance estimation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 963–966.
[97] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and LLMs. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1952–1962.
[98] Gregory T. Smith. 2005. On construct validity: Issues of method and measurement. *Psychological Assessment* 17, 4 (2005).
[99] Ian Soboroff. 2024. Don't use LLMs to make relevance judgments. arXiv:2409.15133v1.
[100] Seth E. Spielman, Joseph Tuccillo, David C. Folch, Amy Schweikert, Rebecca Davies, Nathan Wood, and Eric Tate. 2020. Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index. *Natural Hazards* 100 (2020), 417–436.
[101] Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. 2021. Deep learning for recommender systems: A Netflix case study. *AI Magazine* 42, 3 (Nov. 2021), 7–18.

[102] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. arxiv:202405.01724v1.

[103] Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the International Conference on Intelligent User Interfaces*. 952–966.

[104] Lynda Tamine and Cecile Chouquet. 2017. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing and Management* 53, 2 (2017), 332–350.

[105] Paul Thomas, Gabriella Kazai, Nick Craswell, and Seth Spielman. 2024. What matters in a measure? A perspective from large-scale search evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 282–292.

[106] Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The crowd is made of people: Observations from large-scale crowd labelling. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 25–35.

[107] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1930–1940.

[108] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–18.

[109] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. arXiv:2406.06519 [cs.IR]

[110] Shivani Upadhyay, Nandan Thakur, Ronak Pradeep, Nick Craswell, Daniel Campos, and Jimmy Lin. 2026. Overview of the TREC 2025 Retrieval Augmented Generation (RAG) Track. In *Proceedings of the Text REtrieval Conference*.

[111] Ellen M. Voorhees. 2025. A journey of language tasks evaluation: A keynote at SIGIR 2024. *SIGIR Forum* 58, 2 (March 2025), 1–10.

[112] Ellen M. Voorhees and Donna K. Harman (Eds.). 2005. *TREC: Experiment and evaluation in information retrieval*. MIT Press.

[113] Ellen M. Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can old TREC collections reliably evaluate modern neural retrieval models? arXiv:2201.11086 [cs.IR] https://arxiv.org/abs/2201.11086

[114] Miriam Wanner, Leif Azzopardi, Paul Thomas, Soham Dan, Benjamin Van Durme, and Nick Craswell. 2025. All claims are equal, but some claims are more equal than others: Importance-sensitive factuality evaluation of LLM generations. arXiv:2510.07083 [cs.CL] https://arxiv.org/abs/2510.07083

[115] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, Article 20 (Nov. 2010).

[116] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

[117] Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, user models, and satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 654–662.

[118] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 587–594.

[119] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the International Conference on World Wide Web*. 1011–1018.

[120] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2025. Overview of the TREC 2025 DRAGUN track: Detection, retrieval, and augmented generation for understanding news. In *Proceedings of the Text REtrieval Conference*.

[121] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 379–388.

[122] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are Large Language Models Good at Utility Judgments?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1941–1951.

[123] Wenbo Zhang, Fan Zhang, Jia Chen, and Wei Lu. 2025. Towards Better Evaluating Multi-Query Sessions: A Measure Based on the Theory of Planned Behavior. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 148–158.

[124] Yuye Zhang, Laurence A. F. Park, and Alistair Moffat. 2010. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval Journal* 13 (2010), 46–69.

[125] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

[126] Justin Zobel. 2022. When measurement misleads: The limits of batch assessment of retrieval systems. *ACM SIGIR Forum* 56, 1 (June 2022).