Evaluating Generative Ad Hoc Information Retrieval

Lukas Gienapp*
Leipzig University and ScaDS.AI
Leipzig, Germany

Harrisen Scells
Leipzig University
Leipzig, Germany

Niklas Deckers
Leipzig University and ScaDS.AI
Leipzig, Germany

Janek Bevendorff
Leipzig University
Leipzig, Germany

Shuai Wang
The University of Queensland
Brisbane, Australia

Johannes Kiesel
Bauhaus-Universität Weimar
Weimar, Germany

Shahbaz Syed
Leipzig University
Leipzig, Germany

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany

Guido Zuccon
The University of Queensland
Brisbane, Australia

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany

Matthias Hagen
Friedrich-Schiller-Universität Jena
Jena, Germany

Martin Potthast*
University of Kassel, hessian.AI, ScaDS.AI
Kassel, Germany

ABSTRACT

Recent advances in large language models have enabled the development of viable generative retrieval systems. Instead of a traditional document ranking, generative retrieval systems often directly return a grounded generated text as a response to a query. Quantifying the utility of the textual responses is essential for appropriately evaluating such generative ad hoc retrieval. Yet, the established evaluation methodology for ranking-based ad hoc retrieval is not suited for the reliable and reproducible evaluation of generated responses. To lay a foundation for developing new evaluation methods for generative retrieval systems, we survey the relevant literature from the fields of information retrieval and natural language processing, identify search tasks and system architectures in generative retrieval, develop a new user model, and study its operationalization.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Language models**.

KEYWORDS

Generative information retrieval, Evaluation, Ad hoc search

ACM Reference Format:
Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3626772.3657849

*Corr. Auth. lukas.gienapp@uni-leipzig.de / martin.potthast@uni-kassel.de

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657849

Figure 1: A search engine results page (SERP) has traditionally been a list of document references (list SERP, left). Many generative retrieval systems now have "reinvented" SERPs as generated texts with references (text SERP, right).

1 INTRODUCTION

The development of large language models (LLMs) has prompted search engines to innovate the way results are presented: using LLMs to directly generate a textual response from a query's results. While LLMs can generate unreliable information [3, 53, 63], conditioning their inference o n relevant search results has emerged as a potential technique to ground generated statements [66, 84]. As textual answers can relieve users of the (cognitive) effort of collecting the needed information from individual search results themselves, the design of some search engine's results pages (SERPs) has changed (Figure 1): instead of the proverbial list of "ten blue links" (list SERP, left), a generated text with references is shown (text SERP, right). The first public prototypes of this kind were You.com's You Chat and Neeva AI, closely followed by Microsoft's Bing Copilot, Google's Gemini, Perplexity.ai, Baidu's Ernie,[1] and other research prototypes [62, 140]. Far ahead of this development, already in 2011 Sakai et al. [109] raised an important question: how can text SERP-based search engines be evaluated? An answer was and is not that straightforward, since the modern theory and

[1]See https://chat.you.com; Neeva has shut down; https://chat.bing.com; https://gemini.google.com; https://perplexity.ai; https://yiyan.baidu.com.

practice of retrieval evaluation is premised on the assumption that search results are presented as list SERPs.[2]

According to list SERP user models, a ranked list of results triggers a certain user behavior like reading the results in order until the information need is satisfied or the search is abandoned. In decades of research, a comprehensive theoretical framework of reliable and validated evaluation methods has been built to assess the quality of result rankings with respect to information needs. Replacing ranked results by a generated text undermines this foundation.

In this paper, we focus on questions related to transferring established list SERP evaluation methodology to text SERPs. Our approach is theory-driven and based on a systematic analysis of relevant literature from information retrieval (IR) and related fields. Our contributions relate to the system, user, and evaluation perspectives. Starting with a definition of what generative ad hoc retrieval is, we distinguish two fundamental system models for generative retrieval and contextualize them in Broder's [14] taxonomy of search tasks (Section 2). We then devise a user model for text SERPs, grounded in related behavioral studies (Section 3). Finally, we revisit IR evaluation methodologies to develop a foundation for text SERP effectiveness measures and for the reliable evaluation of generative ad hoc retrieval (Section 4).

2 GENERATIVE RETRIEVAL

In this section, we define the task of generative ad hoc retrieval, we review the two fundamental paradigms of its operationalization, we discuss its contribution on top of traditional ad hoc retrieval, and we distinguish it from other generative retrieval tasks.

2.1 Generative Ad Hoc Retrieval

Ad hoc retrieval refers to scenarios where a user submits a single query and expects the underlying information need to be satisfied by a single result set (i.e., the information need must be satisfied without knowing any previous queries or interactions). At first glance, this ad hoc retrieval task and the task of language generation seem to be quite different. However, retrieval systems and generative language models are both built using document collections D (see Figure 2, top), and the usefulness of both depends on tuning them with user needs, expressed as queries or prompts Q. Users of a retrieval system want to retrieve the most relevant documents for a query, and users of a generative language model want it to generate the most helpful text for a prompt. From an IR perspective, the most salient difference is that a retrieval model ρ induces a ranking on a *finite* document collection D, while a generative language model ψ induces a ranking on the *infinite* set of all possible texts $\mathcal{T}$; generative models were thus recently also framed as infinite indexes [32]. In practice, though, retrieval models only return the top-k results, and generative language models only return one of the possibly many relevant texts from $\mathcal{T}$.

As a retrieval model ρ can only return existing documents, the information available in the underlying collection D determines the degree to which a user's information need can be satisfied. Still, the user has to examine the returned documents for the desired

[2]Even though research on search interfaces has suggested and studied many interaction designs and variants of result presentation [49, 71, 131], with the growth of the Web, the list SERP design became a de facto standard for web search.

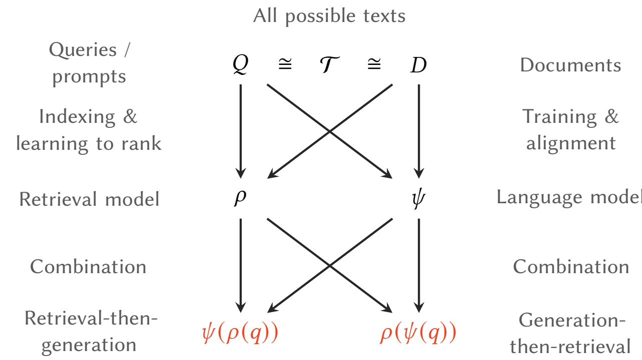

Figure 2: In generative ad hoc retrieval, a retrieval model is combined with a language model. The notation assumes ρ and ψ have texts from $\mathcal{T}$ as input and output, and that they can be complex pieces of software, like Google or ChatGPT.

information. A generative language model ψ instead attempts to alleviate the effort of examining documents by returning a tailored response that integrates all desired information. Yet, the factual accuracy of current generative language models is often prone to confabulations or hallucinations [3, 53, 63, 144] (i.e., there is only a very small subset of accurate texts among all possible texts $\mathcal{T}$).[3]

The term 'generative ad hoc retrieval' refers to approaches that combine the advantages of retrieval and generation in ad hoc scenarios (one query, one result) by retrieving relevant documents from D and generating an answer from them, or by generating a response and "verifying" its statements by retrieving supporting documents from D (Figure 2, bottom left resp. right).

2.2 Two Operationalization Paradigms

Systems for generative ad hoc retrieval require a retrieval component to gather existing documents from a collection for a query, and a generation component to generate a text for a prompt. These components can be combined following two different paradigms [42]: *retrieval-then-generation* or *generation-then-retrieval* (Figure 2, bottom). In a retrieval-then-generation approach, a language model is conditioned with retrieved source material, for instance, by adding evidence to its input prompt [51, 59, 65, 118], by attending to retrieved sources during inference [13, 45, 66], by chaining ideas [54], or by iterative self-attention [143]. In a generation-then-retrieval approach, the retrieval model is used to find sources for generated text passages. Though this idea has received less attention so far [7], it resembles retrieving references for individual generated statements, similar to claim verification [127].

With increasing inference speeds of generative language models, arbitrarily ordered combinations of multiple retrieval and generation steps are possible, leading to *multi-step generative ad hoc retrieval*. The simplest form might be iterative cycles like generating a text passage that is used as a query to retrieve relevant sources, which in turn serve as context for the next generation, etc.

[3]For counterfactual information needs (e.g., What if Columbus didn't discover America? [60]), strong confabulation capabilities could be explicitly desirable, though.

Table 1: Top rows: Broder's (2002) identified generations of web search systems (Gen.) and the tasks from his taxonomy [14] that each generation additionally supports (+). Bottom row: Generative retrieval systems constitute a new 4th generation that aids users in "synthetic" search tasks that require a system to synthesize and condense information.

Gen.	Search task	Information source	User intent	Year
1^{st}	informational	Document	Acquire	1995
2^{nd}	+ navigational	+ Document relations	+ Reach	1998
3^{rd}	+ transactional	+ Search verticals	+ Perform	2002
4^{th}	+ synthetic	+ Generative models	+ Condense	2023

Applications are the continuous generation of text [55, 102, 116], retrieving sources in multiple steps [97], or the refinement of a text through iterative inference [7, 58].

In this paper, we focus on the evaluation of the text SERP output of (possibly multi-step) generative ad hoc retrieval, but we do not consider evaluating any step individually.

2.3 Generative (Ad Hoc) Search Tasks

In 2002, Broder suggested a now well-known taxonomy of search tasks [14] and related them to three generations of web search systems (see Table 1). Each generation utilizes a new source of information in addition to those of its predecessors to meet new user intents. First-generation systems support informational tasks, relying only on the information found within some single document to support a user's intent to acquire (parts of) that information. Second-generation systems additionally exploit document relations, supporting users to reach a specific site, document, or the most authoritative one among many alternatives (i.e., navigational tasks). Third-generation systems blend results from different possibly multimodal vertical systems into a single SERP to support a user in performing transactional tasks.

We argue that generative retrieval systems can be seen as a new fourth generation of web search systems. Their synthesis of a single result "document" that condenses information from different sources relevant to some information need promises to reduce the users' cognitive load compared to prior system generations that required users to condense the information themselves.[4] Additionally, the "synthesizing" nature of generative retrieval systems can conceivably be exploited to generate new pieces of information not contained in the retrieved sources, rendering the generative model itself another new source of information.

While many of the search tasks addressed by generative retrieval systems may seem to be informational in nature, we still suggest to also separate the search tasks in a new category of *synthetic search tasks*. Complex needs like argumentative questions (Should society invest in renewable energy?) or decision-making questions (Should I get life insurance?) are simply not represented that well in Broder's original categories. In contrast to informational tasks, the required information is hardly contained in some single document but rather spread across multiple documents; in contrast to navigational tasks, no single page is anticipated by the user to be reached; and in contrast to transactional tasks, the information condensation should be performed on the retrieval system side but not on the user side. Interestingly, as if already foreseeing generative retrieval, Broder even explicitly constrained informational queries and first-generation systems to static content: "The purpose of such [informational] queries is to find information assumed to be available on the Web in a *static form*. No further interaction is predicted, except reading. By static form we mean that the *target document is not created* in response to the user query." [14, page 5].

The new fourth generation of web search systems supports synthetic search tasks and enables users to access a single, comprehensive generated answer document that can cover in-depth analyses of multiple perspectives on some complex information need. Although the Web may actually also offer some set of documents to satisfy complex needs, an ideal generative retrieval system can directly *dynamically* address them by retrieving relevant documents, synthesizing missing information, and condensing a coherent answer grounded in the retrieved sources.

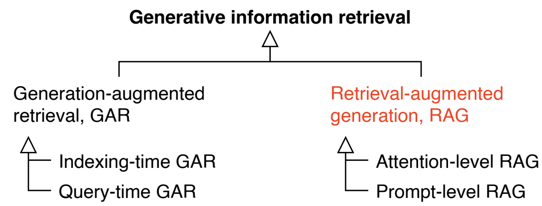

Figure 3: Taxonomy of generative information retrieval and its two main instantiations: generation-augmented retrieval (GAR, yielding list SERPs) and retrieval-augmented generation (RAG, yielding text SERPs; focus of this paper).

2.4 A Taxonomy of Generative Retrieval

'Generative retrieval' or 'generative IR' are umbrella terms for a diversity of approaches that use generative models to solve retrieval tasks.[5] Following Arora et al. [6], Figure 3 categorizes these approaches into generation-augmented retrieval (GAR) and retrieval-augmented generation (RAG). Notably, GAR approaches create traditional list SERPs, while RAG approaches generate text SERPs.

In GAR approaches, generative models are used to enhance the traditional search architecture at indexing time or at query time. At indexing time, generative models can be used for augmenting documents [37, 44, 78, 94, 152] with confabulated or hallucinated content, or for replacing the standard indexing process with what are commonly termed 'differentiable indices' by, for instance, generating document identifiers like page titles [20, 31, 125], URLs [153], or (structured) string identifiers [124, 128, 148, 151]. At query time, generative models can be used for augmenting queries [2, 77], or for modeling relevance by, for instance, generating parts of existing documents from the query and retrieving the documents by string matching [11], by predicting a (re-)ranking directly [123], or by using special tokens as relevance signal [76, 93, 99, 150].

[4] Sakai et al. [109] had proposed to present lists of short automatically identified relevant information nuggets instead of complete documents in 2011, but they had not considered the aspect of condensing the nuggets to a single result.

[5] See also the recent SIGIR workshop on generative IR [10].

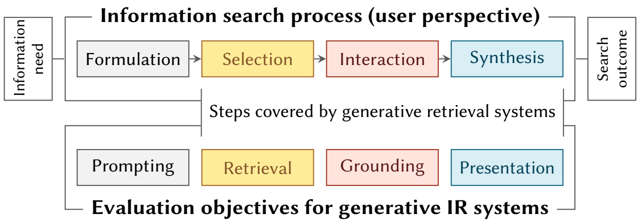

Figure 4: The information search process [126] transforms an information need into a search outcome (top row). Respective corresponding evaluation objectives allow the derivation of a user model for an evaluation setting. Generative IR systems cover the steps of 'selection', 'interaction', and 'synthesis', for which we formulate the corresponding evaluation objectives 'retrieval', 'grounding', and 'presentation' (bottom row).

In RAG approaches—the focus of our paper—, generative models are augmented with retrieval capabilities; either internally as 'attention-level RAG', where the context attended to during generation is retrieved concurrently [13, 45, 54], or externally as 'prompt-level RAG', where the retrieved context is inserted into the prompt. Orthogonally, one can distinguish the RAG variants retrieval-then-generation and generation-then-retrieval (cf. Section 2.2). Beyond GAR and RAG, generative models can also be used to directly generate a response without relying on retrieved information [106], i.e., as infinite indexes [32]. This may involve generating multiple candidates and selecting the best one or regenerating a new response conditioned on the previous ones [135]. Moreover, an answer to a generative ad hoc request can also be the first turn of a conversational search [27, 101, 111], where generative models have led to new tools [85, 141] and dialog options [138].

3 A USER MODEL FOR GENERATIVE IR

The general structure of an information search process [126] as seen from the users' perspective is shown in Figure 4 (top row). After formulating an information need and selecting and interacting with some search results, in a final synthesis step the users try to reach a satisfying outcome. While traditional list SERPs mainly assist the users during selection and interaction, the text SERPs of generative systems also directly encompass the synthesis step. Evaluating retrieval systems with respect to the information search process often relies on some model of user expectations and behavior. Yet, most current user models focus on list SERPs but not text SERPs. Thus, after preliminary considerations (Section 3.1), we explore how the information search process relates to generative approaches (Section 3.2). Afterwards, we follow the evaluation methodology proposed by Agosti et al. [1]: We define generative IR-oriented evaluation objectives for the search process (Section 3.3; shown in the bottom row of Figure 4) and we devise a user model corresponding to these objectives (Section 3.4). In Section 4, we then operationalize the user model.

3.1 Preliminary Considerations

Evaluation Setting. Traditional search results (list SERPs) are ranked lists of documents, each typically referenced by a linked title, snippet, and URL. In generative IR, instead, the search result is a textual response (text SERP), i.e., a sequence of statements, each optionally referenced to sources of evidence. A statement can be any consecutive passage of text, ranging from phrases to sentences or even longer paragraphs. In this context, we consider statements as atomic in the sense that we disregard the nesting of statements of different lengths, and in the sense that statements support claims that are pertinent to the user's information need—comparable to the concept of 'atomic/semantic content units' [74, 91] in summarization evaluation, or 'information nuggets' / 'retrieval units' in traditional IR [21, 29, 108, 109]. A statement can be referenced to none, one, or more sources in form of explicit links to web documents containing the information on which the generated statement is based and by which it is grounded. In this paper, we consider the evaluation to be ad hoc, i.e., based on a single query without search session-based or conversational elements.

Evaluation Paradigms. To estimate the effectiveness of list SERP-based retrieval systems, offline evaluation following the Cranfield paradigm [23] is a de facto standard in IR research. The users' satisfaction with the results for a given topic (query) is estimated by deriving effectiveness scores based on judging a pool of documents returned by the evaluated systems [114]. The pools often are also reused later to evaluate new search systems by checking whether their retrieved results previously were judged—and simplistically assuming non-relevance for the results without previous judgments [39]. However, as the output "documents" of generative retrieval systems may be novel every time, simply assuming non-relevance would not lead to helpful evaluation results. Instead, more sophisticated transfer methods are required to adapt offline evaluation to generative retrieval. Besides offline evaluation, generative retrieval systems could also be evaluated in an online fashion [110]. Online evaluation does not rely on previous judgments but tries to estimate the output of some system by collecting explicit or implicit user feedback [57] like user satisfaction ratings or clicks. This form of evaluation increases the manual effort, often happens in uncontrolled setups, may be expensive and time-consuming to conduct, and is challenging to replicate, repeat, and reproduce [104]. To mitigate these issues especially in an academic setting with limited access to human user data, some studies suggested user simulation to analyze (interactive) information systems [16, 80, 81, 139]. However, simulated users cannot yet compete with "real" human feedback. Recently, fully automatic evaluations, where the output of one system is judged by another, has been proposed as a possible way forward [72, 137]. But judging the output of generative models by means of other models has itself already been criticized [9, 36, 108].

3.2 Steps of the Information Search Process

To derive suitable evaluation objectives for generative ad hoc retrieval, we consider the general user side search process for which Vakkari [126] has suggested to differentiate four steps: search formulation, source selection, source interaction, and information synthesis (Figure 4, top row). Interestingly, each of these steps can be mapped to capabilities of generative retrieval systems.

First, during formulation, the user comes up with a specific query that expresses their information need. This is no different in generative retrieval systems, though what is called a 'query' in IR is often called a 'prompt' for generative systems. To avoid confusion, we stick to the term 'query'. Still, in this paper, we leave the formulation step entirely to the user who may iteratively adapt their search formulation. Yet, we do acknowledge that formulation may also be framed as a system task with the goal of enhancing the users' original query with more context or prompt templates, akin to query suggestion and query expansion in traditional retrieval. Second, during selection, traditionally, the user is presented with a result list possibly containing surrogates like snippets that help to assess whether some result aligns with the user's information need and should be selected for further inspection. In generative retrieval, the selection step corresponds to the system selecting sources that contain potentially relevant information. Third, during interaction, traditionally, the user analyzes the content of the selected results more deeply to extract and structure the relevant information that addresses the knowledge gap underlying the user's information need. In generative retrieval, this step also rather is on the system side by, for instance, attending the generation to previously retrieved pieces of information. Finally, during synthesis, traditionally, the user assembles the search outcome by combining relevant information from their interacted sources. In generative retrieval, synthesis corresponds to the model's inference and generation of the response text from the selected sources. Just like for human users, interaction and synthesis may commence concurrently.

3.3 Evaluation Objectives

For each step of the search process, we define a corresponding generative retrieval-oriented evaluation objective. The objectives are not meant as evaluation steps, but rather as potential targets when evaluating a generative retrieval system as a whole.

Prompting Objective. Corresponding to formulation are evaluation aspects related to a model's input prompt like preciseness (Does the prompt target the specific desired outcome?), ambiguity (Is the prompt unambiguous, targeting only the desired outcome?), or contextuality (Does the prompt provide sufficient context to delineate the information need?). While formulation is an important step to evaluate, it is out of the scope of our paper, as the formulation step in our setting is left to the user and as there already is extense work on prompt engineering [41, 73, 105, 119, 122, 129, 133].

Retrieval Objective. Corresponding to selection are evaluation aspects related to the retrieved sources from which a generative system draws its information. These sources (but also any relevant information that was not retrieved) directly impact the quality of the generated response. Therefore, the retrieval objective covers the assessment of a system's ability to identify relevant (aligning with the users' information need), diverse (covering a variety of information), informative (containing valuable information), and correct (providing accurate information) sources from a collection.

Grounding Objective. Corresponding to interaction are evaluation aspects related to a generative retrieval model's ability to attend to source documents as evidence in response generation. Yet, such grounded text generation may suffer from confabulations / hallucinations of broadly two types [83]: intrinsic confabulations (the model wrongly modifies information from the sources) and extrinsic confabulations (the model generates information not present in the sources). As both types can negatively impact the quality of a generated response [75, 83], the grounding objective covers the assessment of a system's ability to correlate its generated output with information from source documents. This includes the ability to identify relevant information in the sources, to paraphrase information (restate some information correctly), and to establish consistency (not produce contradictions to other sources).

Presentation Objective. Corresponding to synthesis are evaluation aspects related to a model's ability to condense relevant information from multiple sources into a single answer. Resembling multi-document summarization, the presentation objective covers the assessment of an answer's conciseness (at a level of granularity sensible given the topic and user [28]), coherence (uniform writing style in the answer), and accessibility (written in an understandable way; again, dependent on the user).

3.4 Components of the User Model

Developing a user model for generative IR is challenging. The traditional user models were focusing on list SERPs and might thus not apply to text SERPs, as, for instance, the search process steps of selection and interaction are undertaken by the system instead. Additionally, little to no user behavior data on text SERPs are available in the academic context (e.g., A/B tests or laboratory studies) to base model validation or development on. To contribute a user model for generative IR, we thus extrapolate from established evaluation practices in IR and related fields like question answering or summarization. We follow the considerations of Carterette [19], who argues that a user model in IR should include three distinct (sub-)models: (1) a utility model that induces a gain function by capturing how each result provides utility to a user, (2) a browsing model that induces a discount function by capturing how a user interacts with the results, and (3) an accumulation model that combines the individual gain and discount values by capturing how individual utility is aggregated.

3.4.1 *A Utility Model for Generative IR.*

Surveying the literature on evaluation in IR and in related fields, we identified ten utility dimensions applicable to generative ad hoc retrieval. Figure 5 shows the dimensions grouped into five categories (coherence, coverage, consistency, correctness, and clarity) with color-coded corresponding evaluation objectives and indicated granularity from which gain is obtained (statement level: from an individual statement in the response; response level: from the response as a whole).

Coherence. Coherence is a response-level dimension of utility referring to the presentation objective and involving the aspects of statement arrangement that should form a narrative without contradictions [100, 117] (i.e., logical coherence: Is the response well-structured?) and of the writing style that should yield readable and engaging responses [18, 56] (i.e., stylistic coherence: Does the response have a uniform style of speech?).

Coverage. Coverage is a response-level dimension of utility referring to the retrieval objective and measuring how well a user's information need is treated by the returned information; it can be

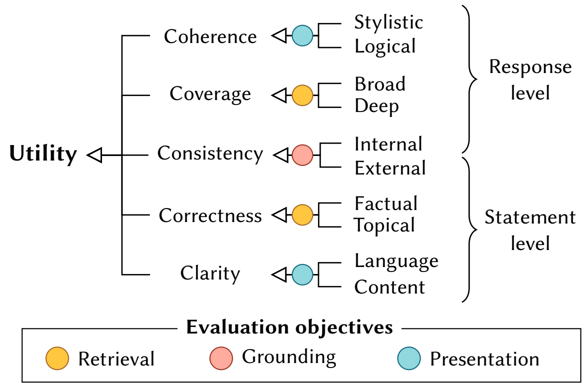

Figure 5: Taxonomy of utility dimensions in generative ad hoc retrieval; colors indicating the evaluation objectives.

subdivided into [17] broad coverage (i.e., whether the response covers diverse information [146]), and deep coverage (i.e., whether the response provides in-depth and highly informative content [82]).

Consistency. Commonly observed problems with source-based text generation are inconsistencies between the sources and parts of the generated text [50] but also inconsistencies between the statements within a response. We refer to the first problem as external consistency, which is a statement-level dimension of utility involving the assessment of the consistency between a statement and its source document(s) to ensure that the generated text aligns in terms of content and context [83, 108, 137] (i.e., Is the statement accurately conveying from the sources?). External inconsistencies are often introduced through model confabulations / hallucinations [53] but they should be distinguished from factual correctness, as external consistency only assesses the alignment of a statement with the sources, and not with some objective truth. To the second consistency problem, we refer to as internal consistency, which is a response-level dimension of utility involving the assessment of the consistency between the responses' individual statements to ensure no contradictions [18, 92, 108]. It should be noted that this does not mean that different conflicting perspectives on a topic can not be reflected in the response, however, these should then be explained. Both notions of consistency refer to the grounding objective.

Correctness. Correctness is a statement-level dimension of utility referring to the retrieval objective and measuring to which degree the information provided in the response is factual, reliable, and addressing the user's information needs. We subdivide correctness into factual and topical correctness. The former captures the degree to which a statement reproduces information that can be assumed as objectively true. Yet, outside of small-scale domain-specific evaluation studies [110] fact-checking remains a challenge [89] and is thus often reduced to a simpler approach, framing it in terms of verifiability [72], not truth. Here, the main requirement is that a piece of information can be attributed to a reliable reference [130, 137] (i.e., Does the statement state things that are verifiable?). Topical correctness captures whether a statement aligns with the user's information need [79, 107, 134] (i.e., Does the statement state things within the scope of the user's information need?).

Clarity. The response of a generative retrieval system should be expressed in a clear and understandable way [112, 149]. This, on the one hand, comprises language clarity: concise [28, 108], comprehensible [17], lexically and grammatically correct, and user-accessible responses. Note that language clarity does not reflect fluency, which is assumed already at human-level for model-generated text [108], but rather the response being in the appropriate language register. For example, a technical query might warrant an academic style of writing in the response, while a joke question might afford a more jovial tone. On the other hand, clarity also comprises content clarity: in order to make a response explainable, the way a statement is written should always clearly communicate the most salient information [115] and where it stems from [95]. Both notions of clarity refer to the presentation objective at the statement level.

3.4.2 A "Browsing" Model for Generative IR. For list SERPs, user interaction is modeled by a set-based or a ranking-based browsing model. In set-based browsing users are assumed to indiscriminately examine all retrieved documents (e.g., for systematic reviews), while in ranking-based browsing users are assumed to traverse the retrieved documents by rank, stopping when either their information need is fulfilled or the search is aborted [19] (e.g., web search). Aborting a search is usually motivated by the information need being satisfied or the effort being too high to justify continuing to browse. Yet, in generative IR, the selection and interaction steps of the search process are undertaken by the system, so that the user only has to read the (often short) generated text. This reduces the effect of effort-based stopping criteria, with most users only aborting their search when their knowledge gap is fulfilled or when the response is deemed insufficient. This is neither really set-based, as reading the response from the beginning and early stopping might occur, nor traditionally ranking-based, as aborting the search is not motivated by effort but rather by search (dis-)satisfaction.

Instead of a browsing model, we thus propose a *reading* model for generative IR to reflect the attention a user places on the response statements while reading. But as there are no dedicated studies on reading behavior for generative IR yet, we turn to related work on reading behavior for document comprehension. In a literature survey, we identified six characteristics from which we deem three as appropriate for our reading model: progression [15, 68, 69, 132, 147] implies that users parse a document sequentially (i.e., reading the statements in their textual order), decay [38, 68, 69, 132, 147] implies that the reading attention diminishes over the span of a text, and saturation [68, 69] implies that users abort when they have read enough to satisfy their information need.

Besides these three characteristics, we deem three others as superfluous for our proposed reading model. First, perceived relevance may be heightened following a relevant statement [68, 147] but we adopt the same restriction as static browsing models for ad hoc retrieval evaluation and neglect inter-statement effects [86, 88]. Second, although reading attention may be highest around query terms [68, 147], our statement- and response-level utility granularities render per-token effects rather constant. Third, although users may skip non-relevant content during reading [15, 48, 68], we ignore this effect in the reading model as non-relevant statements will receive zero utility anyway.

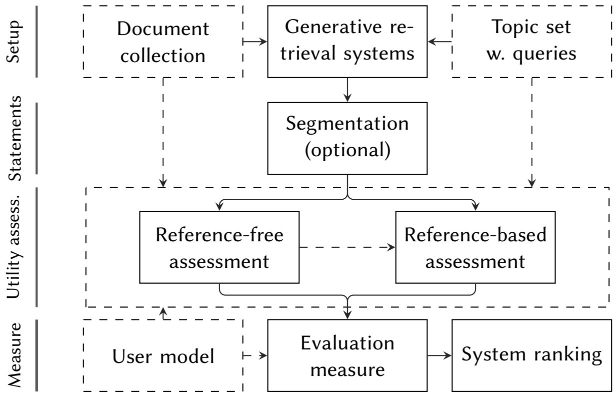

Figure 6: Overview of the evaluation procedure for generative ad hoc retrieval. Given documents and topics, a system produces responses, which are segmented into statements, and assessed for utility, based on which an evaluation measure ranks systems by effectiveness. Solid lines indicate process flow, dashed lines contextual information sources.

Altogether, our proposed reading model thus reflects a sequential reading with decaying attention and early stopping when saturated. These properties can easily be related to the $C/W/L$ framework [87] of browsing models for list SERPs. Sequential reading indicates that the framework's assumption of a sequential process applies, decay (diminishing attention) is related to the framework's conditional continuation probability C and weight W (probability of a user reaching some step of a sequence), and early stopping (saturation) is related to the framework's probability L that indicates whether an item is the last one before aborting a search. Our proposed reading model can thus be operationalized as a monotonically decreasing weight function over statements that discounts the contribution of later statements in a response. At the same time, this directly induces a response structuring approach of putting the most important pieces of information first followed by less important details—similar to the inverted pyramid scheme of news articles [98].

3.4.3 An Accumulation Model for Generative IR. To combine gain and discount values over the statements of a response, we argue in favor of *expected total utility* accumulation [19, 88]. It considers the total utility a searcher accumulates from the whole response. Alternatively, measures could be based around estimating the total 'cost' of accruing information from the response in terms of the effort expended [19]. However, we argue that the effort is comparatively small in text SERPs so that optimizing for it is not that suitable to reliably differentiate systems in evaluation.

4 OPERATIONALIZING EVALUATION

This section considers operationalizations of the proposed user model. The goal is to take stake in what possibilities exist for each step of the process, in an effort to illustrate the required components and how they can be implemented. These considerations are summarized in Figure 6, with each component (rows in the figure) described in a subsection below. The experimental setup encompasses a document collection, a set of topics reflecting the search task, and a set of generative retrieval systems to be evaluated (Section 4.1). Their responses to queries are (optionally) split into statements using a segmentation approach (Section 4.2). Statements are then assessed for their utility, distinguishing between assessment without prior reference, and assessment in relation to prior reference material (Section 4.3). Given annotations and an evaluation measure, the systems can then be ranked with respect to their effectiveness as indicated by an aggregated score (Section 4.4). In each of these four steps, we survey relevant literature and juxtapose proposed evaluation processes with regard to their advantages and disadvantages in the context of the assumed user model.

4.1 Experimental Setting

The established approach for the reproducible evaluation of traditional retrieval systems in an academic context is offline evaluation [23, 114]. It encompasses a document collection, a set of topics reflecting the information needs stated by users, and the set of systems to be tested. Generative retrieval evaluation does not diverge from this basic procedure. Yet, the set of topics should include ones that reflect the search task for which generative retrieval systems are useful, i.e., the synthetic task posited in Section 2.3. Furthermore, a ranking of documents could be pre-supplied for each topic's query in order to exclusively study the systems' synthesizing ability. These can be taken from a baseline retrieval system, shared task results [24–26], or query logs [103]. While opting for offline evaluation allows to reuse established experiment infrastructure such as the TREC format specifications for run and utility judgment files,[6] generative retrieval systems introduce new requirements. Specifically, a run file represents a text SERP, and should thus include the generated text instead of a ranked list of document identifiers. Utility judgments should be persisted together with the annotated text, since no static document identifiers are available.

4.2 Segmenting Statements

While the complete response provided by the system can be annotated as-is (this is especially warranted for response-level utility), in order to ease annotation, it can be segmented into retrieval units (suitable for statement-level utility). This approach of subdividing a response into smaller units is well established in evaluating generated texts in NLP [29, 74, 91], and has been proposed for IR as well [108, 109]. Unit statements should be atomic, in the sense that an assessor should be able to make an informed and reliable decision about their utility with little to no surrounding context.

To this end, human judges can be employed to extract statements [29, 30], but the high effort and low repeatability, as well as the inability to assess the effectiveness of a new system without repeated human intervention renders this approach impractical in most settings. Automatic means of statement segmentation, comparable to the established task of web page segmentation [61], could include splitting after each given reference (useful for experiments investigating grounding, as each statement has a clear attributable source), sentence-level splitting (useful for fine-grained utility dimensions such as correctness or coverage), or prompting the model to output already delineated statements.

[6] https://github.com/usnistgov/trec_eval/

4.3 Assessing Utility

Two different settings for collecting utility assessments can be discerned: (1) a direct assessment of the responses is carried out, without comparing to a separate ground truth; and (2) the unjudged responses can be compared to pre-existing reference responses on the same document and/or query set. The first is similar to reference-free evaluation in summarization [35], which instructs annotators to assess the summary directly, while the second is similar to reference-based evaluation in summarization [12], which instructs annotators to assess the overlap between the system output and reference response, under the assumption that the reference response is the gold standard, or at least exemplary of utility. Not all utility dimensions can be judged on the generated text alone (as, e.g., clarity of language can), but also require information beyond the generated text (e.g., topical/factual correctness). We therefore discern reference responses and context: reference responses are one or more pre-existing texts to which a new response is compared, while context covers the assessment information required. An assessment made with context only is therefore deemed reference-free.

Reference-Free Assessment. To operationalize reference-free evaluation for generative IR, the straightforward approach is to task human judges with assessing a given output. Yet, possibilities also include using the self-reported uncertainty of generative models with out-of-domain data [90], or relying on other generative models to assess the quality of the output, such as BARTScore [136] or GPTScore [40]. Classifiers trained to estimate the magnitude of a utility dimension have also been used [64]. Ranking, either in a pairwise or listwise fashion is an additional form of assessment, i.e., tasking a judge with ordering statements of unknown utility with respect to a given utility dimension [43], under the hypothesis that a response with higher utility will be ranked higher, too.

Reference-Based Assessment. To operationalize reference-based assessment, commonly a similarity measure is applied between reference and response. Lazaridou et al. [65] evaluate their generative retrieval system for the task of question answering by matching words between generated response and the gold answer. Similarity, Arabzadeh et al. [4] assign relevance scores to candidate answers in a QA task by measuring their similarity to annotated ground truth data in latent space. Other content overlap metrics, though not necessarily transferable to the setup proposed here, such as BLEU [96], NIST [33], ROUGE [70] TER [120], METEOR [8], BERT Score [142], or MoverScore [145] have been used to compare a generated text to a reference text, either in full or at the statement level. Ranking models have also proven useful for the relative assessment of generated texts in comparison to references, e.g., in machine translation [34, 121], both in a listwise [67] as well as a pairwise setting [46, 47]. Arabzadeh et al. [4] implement a kind of pseudo-relevance feedback by retrieving candidate reference documents from a corpus, using highly-ranked ones as references.

4.4 Measuring Effectiveness

For statement-level evaluation, the individual utility of statements has to be combined into an overall score for the response. Effectiveness measures for the proposed aggregation model of expected total utility take the general form $\sum_{i=1}^{k} g(d_i) \cdot \sum_{j=i}^{k} p(j)$ [19], where k is the evaluation depth, or in our case, response length, $g(d_i)$ is the utility of the statement at position i, and $p(j)$ is the probability of the user aborting their search immediately after position j. The former is referred to as a gain function, given by the utility assessments of statements collected before. The latter as a discount function, chosen based on prior information about typical user behavior. The widely established measures of DCG and nDCG [52] used for traditional IR evaluation stem from this family of measures [19] and seem suitable for generative retrieval evaluation as well. Yet, they assume a logarithmic discount function. It is currently unclear if this is an appropriate choice to model the effect of decay and saturation in the proposed reading model for generative IR. While the family of measures is thus applicable, the concrete choice of measure needs further empirical validation from user experiments.

For response-level evaluation, two choices for measuring effectiveness exist: either utility is annotated directly for a response, or it is aggregated from individual statement utility. While the latter seems counterintuitive to the response-level vs. statement-level distinction made for utility before, note that the level of granularity on which a utility dimension is defined, and the level of granularity at which annotations are collected can differ. Response-level utility may be aggregated from annotations of individual statements, or statement utility may be derived from annotations of the whole response. For example, consider the response-level utility dimension of broad coverage. It can be estimated by measuring the breadth of topics occurring over all statements, hereby annotating which topics occur in each statement. The previously motivated family of DCG-type measures can be extended to support such evaluation. For example, measure modifications similar to α-nDCG [22] that reward a diverse set of topics in a ranked list can be made for generative IR as well. Independent of how a single score is produced for each response, the final system score is aggregated over multiple topics, increasing robustness and enabling statistical testing.

4.5 Comparison with Existing Frameworks

Two other approaches for the evaluation of generative retrieval systems have been proposed recently: SWAN [108] and EXAM [113]. The starting point of both is a text SERP response, albeit less formalized and without considering the synthetic search task it enables.

SWAN follows a similar approach as is proposed here, first establishing the notion of 'information nuggets', i.e., statements, that constitute the response. Then, a total of 20 categories are described, indicating how a nugget may be scored. The individual nugget scores are then averaged over the whole response. Here, too, two different levels of score categories, i.e., utility dimensions are considered. While similar, our approach and SWAN differ in three important aspects. First, we base our method on a theoretical foundation in the form of a user model, whereas SWAN is mainly motivated from a standpoint of practicability. Second, SWAN is geared towards conversational search, while we consider the ad hoc search task. And third, the utility dimensions we propose differ from SWAN due to the shift in scope: We exclude dimensions specific to conversational search (e.g., recoverability, engagingness), and also those which do not serve to operationalize evaluation for the synthetic search task specifically (such as non-toxicity, robustness to input variations, etc.). The majority of the remaining utility dimensions from SWAN can be mapped to ours.

EXAM takes a completely different approach. Instead of directly evaluating inherent qualities of the generated text, it considers the downstream effectiveness of a Q&A system that ingests the generated answer on multiple-choice questions. The hypothesis is that the correctness of its responses are correlated with the quality of the generated text it uses as input. Being an automatic evaluation method, this allows for rapid experimentation, yet exhibits three major drawbacks: It offers no fine-grained insight into the quality of the generated text, it is not grounded in a user model, and it requires a suitable Q&A system, impacting reliability and comparability, since there are no accepted standards.

In sum, our approach can be related to existing methods in terms of compatibility, complementarity, and consistency. Our approach is compatible with SWAN, as it is derived from similar assumptions, yet adding a theoretical foundation, and constructed with a different search task in mind. Our approach is complementary to EXAM, as our focus is on fine-grained, reliable, user-oriented evaluation, whereas EXAM excels for rapid, system-oriented experimentation with little overhead. Furthermore, our approach is consistent with traditional IR evaluation techniques, making only small adaptations to the utility, browsing, and aggregation models to accommodate the new search paradigm. We believe that this renders much of the work on methods and theoretical foundation for traditional IR evaluation still applicable.

5 CONCLUSION

Generative retrieval introduces a new paradigm for the retrieval of information. With it comes the need to measure and understand new utility dimensions that make text SERP responses from generative retrieval systems relevant to a user's information need. In this paper, we have extrapolated a theoretical foundation for the evaluation of generative retrieval systems from traditional IR and related disciplines. First, we established that the search task of generative ad hoc retrieval goes beyond acquiring information, and instead enables the condensation of information, a process we dub the 'synthetic search task'. Second, we proposed a new user model that accommodates this task, including evaluation objectives, utility dimensions, and a browsing model for text SERPs. Finally, we outlined how one could operationalize the evaluation of generative retrieval systems, surveying how existing evaluation approaches relate to, and could fit into the proposed methodology.

Many techniques for constructing generative retrieval systems are currently emerging, but evaluating their output is still a non-standardized and thus hardly comparable effort, lacking a theoretical motivation. We have provided our vision of a comprehensive approach for evaluating generative retrieval systems. Yet, we believe that user experiments are needed to effectively apply this theoretical motivation, and studying its reliability and validity. This requires a meta-evaluation, such as recently started by Arabzadeh and Clarke [5], of both, existing measures and measures modified for generative IR specifically, to study how well they align with user preferences, and to study the proposed utility dimensions and their ability to reflect user satisfaction, similar to studies conducted for traditional IR [17]. In addition, investigating user interactions with generative retrieval systems is warranted; for example, are user clicks on cited documents in a generated response indicative of their relevance or the user's disbelief, or will generative retrieval make clicks superfluous?

Limitations. The evaluation process we propose in this paper is limited in two ways. First, we opted for a *holistic* evaluation of text SERPs, i.e., instead of evaluating the pipeline of components that constitute the generative retrieval system individually, we focus on evaluating the final response. Second, the evaluation is additionally limited to answer the question if a generative retrieval system is successful at supporting the synthetic search task. This does not consider the more general evaluation objectives that all search systems are subject to (such as bias, fairness, ethicality, or user privacy). In that sense, our considerations are *specific* to generative IR, disregarding the evaluation of *systemic* aspects of IR as a whole. This is not meant to deemphasize the importance of evaluating, e.g., bias in search results, but rather considers it to be outside the scope of this paper.

ACKNOWLEDGMENTS

This publication has been partially supported by the ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence, funded by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus; by a Research Fellowship for Harrisen Scells from the Alexander von Humboldt Foundation; and by the OpenWebSearch.eu project, funded by the European Union (GA 101070014).

REFERENCES

[1] Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari. 2014. Evaluation Methodologies in Information Retrieval (Dagstuhl Seminar 13441). *Dagstuhl Reports* 3, 10 (2014), 92–126.
[2] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 1869–1873.
[3] Hussam Alkaissi and Samy I. McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 15, 2 (2023), 4 pages.
[4] Negar Arabzadeh, Amin Bigdeli, and Charles L. A. Clarke. 2024. Adapting Standard Retrieval Benchmarks to Evaluate Generated Answers. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14609)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 399–414.
[5] Negar Arabzadeh and Charles L. A. Clarke. 2024. A Comparison of Methods for Evaluating Generative IR. arXiv 2404.04044.
[6] Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. 2023. GAR-meets-RAG Paradigm for Zero-Shot Information Retrieval. arXiv 2310.20158.
[7] AutoGPT Contributors. 2023. AutoGPT: The Heart of the Open-Source Agent Ecosystem. https://github.com/Significant-Gravitas/AutoGPT.
[8] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72.
[9] Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr. 2023. Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. arXiv 2305.01509.
[10] Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-IR @ SIGIR 2023: The First Workshop on Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development*

in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 3460–3463. https://doi.org/10.1145/3539618.3591923

[11] Michele Bevilacqua, Giuseppe Ottaviano, Patrick S. H. Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive Search Engines: Generating Substrings as Document Identifiers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022.* 16 pages.

[12] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 9347–9359.

[13] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *International Conference on Machine Learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 2206–2240.

[14] Andrei Z. Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.

[15] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger van Elst. 2012. Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond. *ACM Trans. Interact. Intell. Syst.* 1, 2 (2012), 9:1–9:30.

[16] Arthur Câmara, David Maxwell, and Claudia Hauff. 2022. Searching, Learning, and Subtopic Ordering: A Simulation-Based Analysis. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 142–156.

[17] Berkant Barla Cambazoglu, Valeria Bolotova-Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and W. Bruce Croft. 2021. Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering. In *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14–19, 2021*, Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith (Eds.). ACM, 75–84.

[18] Robert Capra and Jaime Arguello. 2023. How does AI Chat Change Search Behaviors? arXiv 2307.03826.

[19] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 903–912.

[20] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative Evidence Retrieval for Fact Verification. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 – 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2184–2189.

[21] Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense X Retrieval: What Retrieval Granularity Should We Use? arXiv 2312.06648.

[22] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20–24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 659–666.

[23] Cyril W. Cleverdon. 1967. The Cranfield Tests on Index Language Devices. *Aslib Proceedings* 19, 6 (1967), 173–194.

[24] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16–20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 13 pages.

[25] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Overview of the TREC 2021 Deep Learning Track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15–19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 16 pages.

[26] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2022. Overview of the TREC 2022 Deep Learning Track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15–19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 21 pages.

[27] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34–90.

[28] Hoa Trang Dang. 2005. Overview of DUC 2005. In *DUC 2005, Document Understanding Workshop October 9–10, 2005, Vancouver, B.C., Canada*. 1–12.

[29] Hoa Trang Dang and Jimmy Lin. 2007. Different Structures for Evaluating Answers to Complex Questions: Pyramids Won't Topple, and Neither Will Human Assessors. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23–30, 2007, Prague, Czech Republic*, John Carroll, Antal van den Bosch, and Annie Zaenen (Eds.). The Association for Computational Linguistics, 768–775.

[30] Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 Question Answering Track. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14–17, 2006 (NIST Special Publication, Vol. 500-272)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST), 18 pages.

[31] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net, 20 pages.

[32] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2023. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19–23, 2023*, Jacek Gwizdka and Soo Young Rieh (Eds.). ACM, 172–186.

[33] George R. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research, HLT 2002*. 138–145.

[34] Kevin Duh. 2008. Ranking vs. Regression in Machine Translation Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation, WMT@ACL 2008, Columbus, Ohio, USA, June 19, 2008*, Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron S. Fordyce (Eds.). Association for Computational Linguistics, 191–194.

[35] Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguistics* 9 (2021), 391–409.

[36] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, 39–50.

[37] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. arXiv 2109.10086.

[38] Aline Frey, Gelu Ionescu, Benoit Lemaire, Francisco López-Orozco, Thierry Baccino, and Anne Guérin-Dugué. 2013. Decision-Making in Information Seeking on Texts: An Eye-Fixation-Related Potentials Investigation. *Frontiers in Systems Neuroscience* 7 (2013), 22 pages.

[39] Maik Fröbe, Lukas Gienapp, Martin Potthast, and Matthias Hagen. 2023. Bootstrapped nDCG Estimation in the Presence of Unjudged Documents. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 313–329.

[40] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. arXiv 2302.04166.

[41] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1–6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3816–3830.

[42] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv 2312.10997.

[43] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient Pairwise Annotation of Argument Quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5772–5781.

[44] Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2Query–: When Less is More. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13981)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 414–422.

[45] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv 2002.08909.

[46] Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to Differentiate Better from Worse Translations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 214–220.

[47] Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise Neural Machine Translation Evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26–31, 2015, Beijing, China, Volume 1: Long Papers*. The Association for Computer Linguistics, 805–814.

[48] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-Tracking Measures. In *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26–29, 2014*, David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson (Eds.). ACM, 58–67.

[49] Marti A. Hearst. 2009. *Search User Interfaces*. Cambridge University Press.

[50] Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. 2021. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. arXiv 2104.14839.

[51] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19–23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 874–880.

[52] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[53] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.

[54] Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan, and Graham Neubig. 2022. Retrieval as Attention: End-to-end Learning of Retrieval and Reading within a Single Transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 2336–2349.

[55] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 7969–7992.

[56] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5082–5093.

[57] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1-2 (2009), 1–224.

[58] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP. arXiv 2212.14024.

[59] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net, 69 pages.

[60] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. arXiv 2305.00050.

[61] Johannes Kiesel, Lars Meyer, Florian Kneist, Benno Stein, and Martin Potthast. 2021. An Empirical Comparison of Web Page Segmentation Algorithms. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 62–74.

[62] Bevan Koopman, Ahmed Mourad, Hang Li, Anton van der Vegt, Shengyao Zhuang, Simon Gibson, Yash Dang, David Lawrence, and Guido Zuccon. 2023. AgAsk: An Agent to Help Answer Farmer's Questions from Scientific Documents. *International Journal on Digital Libraries* (2023), 16 pages.

[63] Bevan Koopman and Guido Zuccon. 2023. Dr ChatGPT Tell Me What I Want to Hear: How Different Prompts Impact Health Answer Correctness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 15012–15022. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.928

[64] Alex Kulesza and Stuart M. Shieber. 2004. A Learning Approach to Improving Sentence-Level MT Evaluation. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. 10 pages.

[65] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-Augmented Language Models Through Few-Shot Prompting for Open-Domain Question Answering. arXiv 2203.05115.

[66] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). 16 pages.

[67] Maoxi Li, Aiwen Jiang, and Mingwen Wang. 2013. Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation. In *Proceedings of Machine Translation Summit XIV: Papers, MTSummit 2013, Nice, France, September 2–6, 2013*, Andy Way, Khalil Sima'an, and Mikel L. Forcada (Eds.). 8 pages.

[68] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution during Relevance Judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22–26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 733–742.

[69] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 795–804.

[70] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.

[71] Chang Liu, Ying-Hsang Liu, Jingjing Liu, and Ralf Bierig. 2021. Search Interface Design and Evaluation. *Found. Trends Inf. Retr.* 15, 3-4 (2021), 243–416.

[72] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 7001–7025.

[73] Vivian Liu and Lydia B. Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 – 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 384:1–384:23.

[74] Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 4140–4170.

[75] Klaus-Michael Lux, Maya Sappelli, and Martha A. Larson. 2020. Truth or Error? Towards Systematic Analysis of Factual Errors in Abstractive Summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2020, Online, November 20, 2020*, Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard H. Hovy (Eds.). Association for Computational Linguistics, 1–10.

[76] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-Shot Listwise Document Reranking with a Large Language Model. arXiv 2305.02156.

[77] Sean MacAvaney, Craig Macdonald, Roderick Murray-Smith, and Iadh Ounis. 2021. IntenT5: Search Result Diversification using Causal Language Models. arXiv 2108.04026.

[78] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via Prediction of Importance with Contextualization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1573–1576.

[79] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Syst.* 35, 3 (2017), 19:1–19:32.

[80] David Maxwell and Leif Azzopardi. 2016. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24–28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 731–740.

[81] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19–23, 2015*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 313–322.

[82] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 135–144.

[83] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1906–1919.

[84] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: A Survey. arXiv 2302.07842.

[85] Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017 - System Demonstrations*, Lucia Specia, Matt Post, and Michael Paul (Eds.). Association for Computational Linguistics, 79–84.

[86] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2015. INST: An Adaptive Metric for Information Retrieval Evaluation. In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8–9, 2015*, Laurence Anthony F. Park and Sarvnaz Karimi (Eds.). ACM, 5:1–5:4.

[87] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Trans. Inf. Syst.* 35, 3 (2017), 24:1–24:38.

[88] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users Versus Models: What Observation Tells us about Effectiveness Metrics. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 – November 1, 2013*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi (Eds.). ACM, 659–668.

[89] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mücahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12880)*, K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 264–291.

[90] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. 2019. Do Deep Generative Models Know What They Don't Know?. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net, 19 pages.

[91] Ani Nenkova, Rebecca J. Passonneau, and Kathleen R. McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Trans. Speech Lang. Process.* 4, 2 (2007), 4.

[92] Toru Nishino, Shotaro Misawa, Ryuji Kano, Tomoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2019. Keeping Consistency of Sentence Generation and Document Classification with Multi-Task Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3193–3203.

[93] Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 708–718.

[94] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. arXiv 1904.08375.

[95] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Stevenson, WA, USA, October 28–30, 2019*, Edith Law and Jennifer Wortman Vaughan (Eds.). AAAI Press, 97–105.

[96] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA*. ACL, 311–318.

[97] Quinn Patwardhan and Grace Hui Yang. 2023. Sequencing Matters: A Generate-Retrieve-Generate Model for Building Conversational Agents. arXiv 2311.09513.

[98] Horst Pöttker. 2003. News and its Communicative Quality: The Inverted Pyramid – When and Why did it Appear? *Journalism Studies* 4, 4 (2003), 501–511.

[99] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. arXiv 2306.17563.

[100] Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Comput. Linguistics* 24, 3 (1998), 469–500.

[101] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7–11, 2017*, Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). ACM, 117–126.

[102] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. arXiv 2302.00083.

[103] Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2848–2860.

[104] Gareth Renaud and Leif Azzopardi. 2012. SCAMP: A Tool for Conducting Interactive Information Retrieval Experiments. In *Information Interaction in Context: 2012, IIix'12, Nijmegen, The Netherlands, August 21–24, 2012*, Jaap Kamps, Wessel Kraaij, and Norbert Fuhr (Eds.). ACM, 286–289.

[105] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8–13, 2021, Extended Abstracts*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi (Eds.). ACM, 314:1–314:7.

[106] Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net, 28 pages.

[107] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08–12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 675–684.

[108] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. arXiv 2305.08290.

[109] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. 2011. Click the Search Button and be Happy: Evaluating Direct and Immediate Information Access. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24–28, 2011*, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, 621–630.

[110] Malik Sallam, Nesreen A Salim, B Ala'a, Muna Barakat, Diaa Fayyad, Souheil Hallit, Harapan Harapan, Rabih Hallit, Azmi Mahafzah, and B Ala'a. 2023. ChatGPT Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study at the Outset of a Paradigm Shift in Online Search for Information. *Cureus* 15, 2 (2023), 16 pages.

[111] Gerard Salton. 1969. *Interactive Information Retrieval.* Technical Report. Cornell University.

[112] Mehrnoosh Sameki, Aditya Barua, and Praveen K. Paritosh. 2016. Rigorously Collecting Commonsense Judgments for Complex Question-Answer Content. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8–11, 2015, San Diego, California, USA, Volume 3*, Elizabeth Gerber and Panos Ipeirotis (Eds.). AAAI Press, 26–33.

[113] David P. Sander and Laura Dietz. 2021. EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want. In *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15–18, 2021 (CEUR Workshop Proceedings, Vol. 2950)*, Omar Alonso, Stefano Marchesin, Marc Najork, and Gianmaria Silvello (Eds.). CEUR-WS.org, 136–146.

[114] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends Inf. Retr.* 4, 4 (2010), 247–375.

[115] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-based Explanation Over Text. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21–24, 2022.* ACM, 611–636.

[116] Sina J. Semnani, Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia. arXiv 2305.14292.

[117] Darsh J. Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets Hybrid: Consensual Multi-document Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6–11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 5213–5222.

[118] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv 2301.12652.

[119] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4222–4235.

[120] Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8–12, 2006.* Association for Machine Translation in the Americas, 223–231.

[121] Xingyi Song and Trevor Cohn. 2011. Regression and Ranking based Optimisation for Sentence Level MT Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30–31, 2011*, Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan (Eds.). Association for Computational Linguistics, 123–129.

[122] Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-Theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22–27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 819–862.

[123] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 14918–14937.

[124] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). 13 pages.

[125] James Thorne. 2022. Data-Efficient Autoregressive Document Retrieval for Fact Verification. arXiv 2211.09388.

[126] Pertti Vakkari. 2016. Searching as Learning: A Systematization based on Literature. *J. Inf. Sci.* 42, 1 (2016), 7–18.

[127] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 7534–7550.

[128] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022. A Neural Corpus Indexer for Document Retrieval. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). 15 pages.

[129] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv 2302.11382.

[130] Wikimedia Foundation. 2023. Wikipedia: Verifiability, not Truth. https://web.archive.org/web/20230627143645/https://en.wikipedia.org/wiki/Wikipedia:Verifiability,_not_truth. Accessed: 2023-06-27.

[131] Max L. Wilson. 2011. Interfaces for Information Retrieval. In *Interactive Information Seeking, Behaviour and Retrieval*, Ian Ruthven and Diane Kelly (Eds.). Facet Publishing, 139–170.

[132] Zhijing Wu, Jiaxin Mao, Kedi Xu, Dandan Song, and Heyan Huang. 2023. A Passage-Level Reading Behavior Model for Mobile Search. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 – 4 May 2023*, Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (Eds.). ACM, 3236–3246.

[133] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. arXiv 2309.03409.

[134] Ziying Yang. 2017. Relevance Judgments: Preferences, Scores and Ties. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 1373.

[135] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate Rather Than Retrieve: Large Language Models are Strong Context Generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023.* OpenReview.net, 27 pages.

[136] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27263–27277.

[137] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 4615–4635.

[138] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and Learning from User Interactions for Search Clarification. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1181–1190.

[139] Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. 2022. The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17–21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 4661–4666.

[140] Dake Zhang and Ronak Pradeep. 2023. ReadProbe: A Demo of Retrieval-Enhanced Large Language Models to Support Lateral Reading. arXiv 2306.07875.

[141] Edwin Zhang, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. 2021. Chatty Goose: A Python Framework for Conversational Search. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2521–2525.

[142] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020.* OpenReview.net, 43 pages.

[143] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict Again. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8–12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 418–426.

[144] Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. 2023. Can ChatGPT-like Generative Models Guarantee Factual Accuracy? On the Mistakes of New Generation Search Engines. arXiv 2304.11076.

[145] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 563–578.

[146] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. 2012. Coverage-Based Search Result Diversification. *Information Retrieval* 15, 5 (2012), 433–457.

[147] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human Behavior Inspired Machine Reading Comprehension. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 425–434.

[148] Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, and Ji-Rong Wen. 2023. DynamicRetriever: A Pre-trained Model-based IR System Without an Explicit Index. *Mach. Intell. Res.* 20, 2 (2023), 276–288.

[149] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2009. A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites. In *Proceedings of the 14th International Conference on Information Quality, ICIQ 2009, Hasso Plattner Institute, University of Potsdam, Germany, November 7–8 2009*, Paul L. Bowen, Ahmed K. Elmagarmid, Hubert Österle, and Kai-Uwe Sattler (Eds.). HPI/MIT, 264–265.

[150] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Berdersky. 2023. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. arXiv 2310.14122.

[151] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the Gap Between Indexing and Retrieval for Differentiable Search Index with Query Generation. arXiv 2206.10128.

[152] Shengyao Zhuang and Guido Zuccon. 2021. Fast Passage Re-ranking with Contextualized Exact Term Matching and Efficient Passage Expansion. arXiv 2108.08513.

[153] Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large Language Models are Built-in Autoregressive Search Engines. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9–14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 2666–2678.