Evaluating Generative Ad Hoc Information Retrieval

Lukas Gienapp
Leipzig University and ScaDS.AI

Harrisen Scells
Leipzig University

Niklas Deckers
Leipzig University and ScaDS.AI

Janek Bevendorff
Leipzig University

Shuai Wang
The University of Queensland

Johannes Kiesel
Bauhaus-Universität Weimar

Shahbaz Syed
Leipzig University

Maik Fröbe
Friedrich-Schiller-Universität Jena

Guido Zuccon
The University of Queensland

Benno Stein
Bauhaus-Universität Weimar

Matthias Hagen
Friedrich-Schiller-Universität Jena

Martin Potthast
Leipzig University and ScaDS.AI

ABSTRACT

Recent advances in large language models have enabled the development of viable generative information retrieval systems. A generative retrieval system returns a grounded generated text in response to an information need instead of the traditional document ranking. Quantifying the utility of these types of responses is essential for evaluating generative retrieval systems. As the established evaluation methodology for ranking-based ad hoc retrieval may seem unsuitable for generative retrieval, new approaches for reliable, repeatable, and reproducible experimentation are required. In this paper, we survey the relevant information retrieval and natural language processing literature, identify search tasks and system architectures in generative retrieval, develop a corresponding user model, and study its operationalization. This theoretical analysis provides a foundation and new insights for the evaluation of generative ad hoc retrieval systems.

CCS CONCEPTS

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

KEYWORDS

generative information retrieval, evaluation, ad hoc search

1 INTRODUCTION

The development of large language models (LLMs) has prompted search engine and AI companies to innovate the way search results are presented. LLMs can be used to generate a text that directly satisfies an information need. However, since LLMs can generate unreliable information [2, 54, 155], conditioning their inference on relevant documents has emerged as a potential technique to ground their generated statements [67, 86]. This can relieve users of the (cognitive) effort of acquiring the needed information from individual search results themselves, which affords a change in the design of a search engine results page (SERP; Figure 1): instead of the proverbial list of "ten blue links" (list SERP), a generated text with references is shown (text SERP). The first public prototypes of this kind were You.com's You Chat and Neeva AI, closely followed by Microsoft's Bing Chat, Google's Bard, Perplexity.ai, Baidu's Ernie,[1] and research prototypes [64, 142].

[1]See https://chat.you.com; Neeva has shutdown; https://chat.bing.com (requires the Edge browser); https://bard.google.com; https://perplexity.ai; https://yiyan.baidu.com.

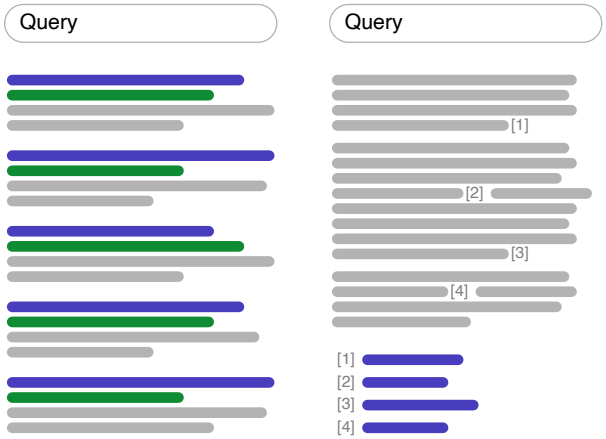

Figure 1: A search engine results page (SERP) has traditionally been a list of document references (list SERP, left). Large language models afford its reinvention as a generated text document with source references (text SERP, right).

Far ahead of this development, Sakai et al. [109] raised an important question: How can search engines that use text SERPs be evaluated? Evaluating text SERPs is not straightforward, since the modern theory and practice of evaluation in information retrieval is built on a user model premised on the assumption that search results are presented as list SERPs.[2] According to this model, the ranked list of documents on a list SERP elicits the user's information behavior, which consists of reading the documents in order until the information need is satisfied or the search is abandoned. In decades of research, a comprehensive theoretical evaluation framework of reliable and validated methods has been built to assess the quality of a document ranking with respect to an information need. Replacing the ranking with a text undermines this foundation.

In this paper, we focus on the basic task of generative ad hoc retrieval and transferring established evaluation methodology for list SERPs to text SERPs. Our approach is theory-driven and based on a systematic analysis of relevant literature from information

[2]Extensive research on search interfaces has included many alternatives, search features, interaction designs, and result visualizations [50, 73, 134]. Nonetheless, with the growth of the web, Google's list SERP design became a de facto standard for web search, and the term "search engine results page" a synonym for "document ranking."

retrieval (IR) and related fields. Our contributions relate to the systems, user, and evaluation perspectives. Starting with a definition of the task of generative ad hoc retrieval, we explore system models for generative retrieval and the tasks they can solve (Section 2). We then devise a user model for text SERPs based on their salient properties and grounded in related behavioral studies (Section 3). Based on both, we transfer established evaluation methodologies from information retrieval as a foundation for new text SERP effectiveness measures (Section 4) and reliable, repeatable evaluation of generative ad hoc information retrieval tasks.

2 THE GENERATIVE RETRIEVAL TASK

In this section, we define the task of generative ad hoc retrieval, review the two fundamental paradigms of its operationalization, discuss its main contribution to traditional ad hoc retrieval, and distinguish it from related generative tasks in IR.

2.1 Generative Ad Hoc Retrieval

Consider the two distinct tasks of retrieval and language generation. As illustrated in Figure 2, IR systems as well as generative language models are created using large collections of documents D. However, their usefulness depends on users' needs and expectations, expressed as a set of queries or prompts Q. The users of an IR system want to retrieve the most relevant documents that satisfy their information needs. Similarly, the users of a generative language model want to generate the most helpful text for their current tasks. From an IR perspective, the fundamental difference between the two is as follows: A retrieval model ρ induces a ranking on a finite document collection D with respect to their relevance to a query q. A language model ψ induces a corresponding ranking on the infinite set of all possible texts $\mathcal{T}$. In practice, the former is used to return the top-k ranked documents from D, and the latter to return, i.e., generate, just one of the many possible relevant documents from $\mathcal{T}$. Generative models like ψ have therefore recently been framed as infinite indexes [31].

Since a retrieval model ρ can only return existing documents, the relevant information (nuggets) in D determines the degree to which a user's information need can be satisfied. The user has to examine the returned documents for the desired information. A generative language model ψ instead attempts to alleviate the effort of examining documents by returning a tailored response that compiles all information required by the user. Yet the factual accuracy of current generative language models is often lacking and prone to hallucinations [2, 54, 146, 155] (i.e., there is only a small subset of accurate documents among all possible texts $\mathcal{T}$). Generative ad hoc retrieval can therefore be described as the task of combining both types of models so that their respective advantages and disadvantages complement or balance each other. For single-shot ad hoc retrieval, two fundamental combination approaches can be distinguished (Figure 2, bottom): retrieval of relevant documents from D based on which a response is generated, or generation of a response and verifying its statements by retrieving supporting documents from D.

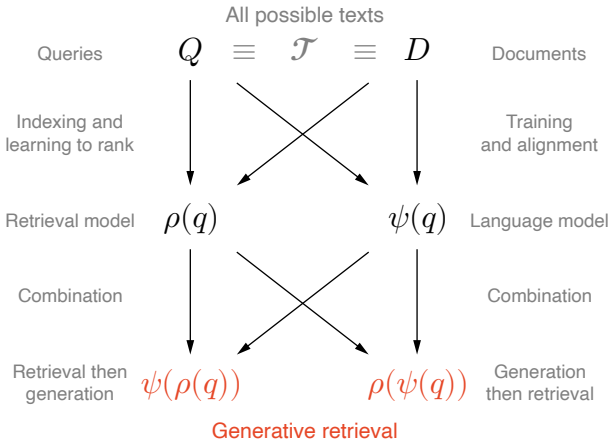

Figure 2: The task of generative ad hoc retrieval entails combining a retrieval model and a language model. The notation waives mathematical rigorousness in favor of intuitive understanding.

2.2 Operationalizing Generative Retrieval

We discern two distinctive components a generative retrieval systems: (1) *retrieval*, where a query is addressed with existing documents from a collection; and (2) *generation*, where a query is addressed by generating a new text. The two fundamental approaches to combining both components, which have also been pursued by existing work include *retrieval-then-generation* and *generation-then-retrieval*. These paradigms can be seen as two atomic approaches to operationalize generative ad hoc retrieval. However, with increasing inference speeds for large language models in particular, and generative AI in general, also combinations of these two paradigms are conceivable, which we refer to as *multi-turn generative retrieval*.

In a retrieval-then-generation approach, a generative process is conditioned with retrieved source material. This can commence by, e.g., adding evidence from retrieved sources to the input prompt of the generative model [52, 60, 66, 118], attending to retrieved sources during generative inference [10, 46, 67], chaining models [55], or iterative self-attention starting from sources [145].

In a generation-then-retrieval approach, instead a retrieval process is prompted with generated text. While this approach has received little attention in existing work [23], it commonly takes the form of retroactively retrieving references for a generated statement, similar to claim verification [131].

In multi-turn generative retrieval, retrieval and generation are combined in an arbitrarily ordered sequence of retrieval and generation steps. This commonly proceeds in a cyclical pattern, where a generated passage is then utilized as a query to retrieve relevant sources, which in turn serve as context for future text generation. This can be employed for continuous generation of text [56, 102, 116], retrieving sources at multiple steps in the process, or for refinement through iterative inference [23, 59]. However, we focus our efforts in this paper solely on the generative ad hoc retrieval task, where we do not consider multiple turns in a conversation.

2.3 Contribution of Generative Retrieval

Generative ad hoc retrieval is a new variant of ad hoc retrieval, the task of satisfying a query's information need independent of all other queries the user or other users submit before or after. Ad hoc retrieval has a long history and large body of research dedicated to it [81]. This raises the question of what generative ad hoc retrieval contributes to traditional ad hoc retrieval.

In this regard, we refer to Broder's taxonomy of web search [11], as compiled in Table 1. It spans three well-known categories of search tasks, and juxtaposes them with three corresponding generations of search engines. Each generation utilizes a new source of information in addition to those of its predecessors to meet new user intents. The first generation of web search engines supports informational tasks, relying on the information found within a document in order to support a user's intent to acquire (parts of) that information. The second generation additionally exploits document relations, supporting users that intend to reach a specific site or document, or the most authoritative one among many alternatives, i.e., information needs that are navigational in nature. The third generation blends results from different vertical search engines, integrating multimedia and multimodal results into a single SERP to support a user in performing tasks.

Generative retrieval systems can be seen as a new, 4th generation of web search engines. They enable the synthesis of new documents relevant and tailored to a user's information need. Given a sufficiently complex information need (i.e., one that cannot be answered by information from a single document), this capability is primarily used to operationalize a user's intent to collect and compile a comprehensive overview of the information required to solve their task, condensed into a long-form text. This part of information behavior, the condensation of information from multiple sources, has previously not been supported to the best of our knowledge. Users therefore had to browse and parse the information from retrieved documents on a list SERP themselves to satisfy their information needs. Generative models relieve users from this extra work and cognitive load, so that they now only have to read and understand a generated text.[3] Additionally, the synthetical nature of such systems can conceivably be harnessed to generate new pieces of information not contained in retrieved sources, rendering the generative model itself a source of information.

While this could be framed as an extension to the informational search task, we argue that it deserves to be treated on its own merits, and therefore postulate the *synthetical search task*. Consider opinionated information needs ("Should society invest in renewable energy?") or decision-making ones ("Should I get life insurance?"). These are not fully supported by the first three generations, since (1) in contrast to informational tasks, information is likely spread across multiple documents; (2) in contrast to navigational tasks, no single page is premeditated to be reached by the user; and (3) in contrast to transactional tasks, the goal, i.e., condensing the information is to be addressed on the system side. Additionally, Broder explicitly constrains informational queries and first generation search 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." [11, page 5]

[3]Sakai et al. [109] has previously proposed to automatically identify relevant information nuggets in retrieved documents and present them in a list, but did not consider the aspect of condensing them.

Table 1: Ad hoc web search system generations (Gen.), and what each supports in addition to (+) the previous one according to Broder [11]. Generative retrieval systems constitute the 4th generation which aids users in synthetical tasks by condensing information using generative models.

Gen.	Search Task	Information Source	User Intent	Year
1st	informational	Document	Acquire	1995
2nd	+ navigational	+ Document relations	+ Reach	1998
3rd	+ transactional	+ Search verticals	+ Perform	2002
4th	+ synthetical	+ Generative models	+ Condense	2023

The fourth generation of search engines supports the synthetical search task and ideally enables users to access a single, comprehensive document that covers a complex topic with in-depth analysis from varied perspectives. Although the web may offer the right (set of) document(s) to answer a such query, the system compiles them, synthesizes missing information, presents it coherently, and is grounding its claims in the retrieved sources.

2.4 Other Kinds of Generative Retrieval

"Generative IR" is an umbrella term to describe a diversity of approaches that combine retrieval and generative components to solve a task.[4] For example, generative models can be augmented with retrieval capabilities or used in an IR pipeline, such as with retrieval-augmented language models [10, 46, 55] or infinite indexes [31]. Furthermore, generative models can be used to enhance a retrieval process [3] by augmenting documents [36, 44, 79, 96, 153] or queries [41, 78] with hallucinated content. The entire retrieval pipeline can also be approached end-to-end by, e.g., generating document identifiers, such as page titles [15, 19, 125], URLs [154], and (structured) string identifiers [124, 132, 150, 152]. Instead of generating identifiers, generating parts of existing documents and performing retrieval by string matching [8] can be highly effective, and a (re-)ranking can also be predicted directly [123].

Generative models can also be used to directly generate a response without relying on retrieved information [110]. This extends to generating multiple candidates and choosing the best or regenerating a new response conditioned on the previous ones [138]. Yet, generative ad hoc retrieval exceeds that by requiring grounding.

Finally, ad hoc generative retrieval is strongly related to, and borrows from several pre-existing fields. Conversational search [27, 101, 111] has led to developing new tools [87, 143], resources [94, 127], and dialogue options [61, 63, 129, 130, 141]. Question answering has been approached with LLMs to produce direct answers [106]. Text summarization [45, 109] has been used in an IR context to, for example, generate snippets [4, 20, 126]. Generative ad hoc retrieval is different from these related tasks as it is broader in scope than question answering systems [71], requires explicit grounding [18], is not interactive like conversational search, and has more information processing requirements than summarization.

[4]See also the recent SIGIR workshop on generative IR [7].

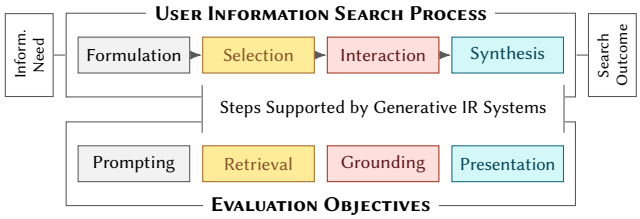

Figure 3: The user information search process [128] transforms an information need into the search outcome. The evaluation objectives allow to derive a user model for an evaluation setting. Generative IR systems span the user steps of *Selection*, *Interaction*, and *Synthesis*, resulting in corresponding objectives *Retrieval*, *Grounding*, and *Presentation*.

3 A USER MODEL FOR GENERATIVE IR

Any IR system should align with user expectations, thus, evaluation needs to be grounded in a user model. Yet, existing user models that have been derived to facilitate evaluation in IR are based on the assumptions of list SERPs. After preliminary considerations (Section 3.1) to derive a text SERP user model, we first consider the general search process of a user (Section 3.2) and explore how it relates to generative approaches. Then, we follow the evaluation methodology proposed by Agosti et al. [1]: first, define evaluation objectives (Section 3.3) and then devise a user model that corresponds to these objectives (Section 3.4). This makes it possible to later derive metrics that operationalize the user model, aggregated over multiple results or queries.

This structure is also reflected in Figure 3. We base our proposed evaluation methodology on the ad hoc information search process [128] as seen from the users' perspective (top of the figure), and formulate evaluation objectives that correspond to each component (bottom of the figure), which take into account the evaluation setting from which a user model can be induced. Traditional IR can assist the user only during Selection with a list SERP. Meanwhile, generative retrieval encompasses all three steps of Selection, Interaction, and Synthesis, to support the synthetical search task and respond with a text SERP. An evaluation of a generative retrieval system should therefore focus on these steps, which are mirrored in the Retrieval, Grounding, and Presentation evaluation objectives.

3.1 Preliminary Considerations

Evaluation Setting. In traditional retrieval, the user is presented with a ranked list of documents (list SERP), each typically referenced by a linked title, snippet, and URL. In generative IR, instead, a response text is presented (text SERP), i.e., a sequence of statements, each optionally referencing one or more sources of evidence in support of the statement. A statement can be any consecutive passage of text, ranging from a single word or phrase to a sentence and even one or more paragraphs. In this context, statements are considered 'atomic' in the sense that we disregard the nesting of statements of different lengths, and that they support one or more claims that are pertinent to the user's information need. They are comparable to the concept of 'atomic/semantic content units' [76, 93] in summarization evaluation, or 'information nuggets' in traditional IR [29, 108, 109]. A statement can be referencing none, one, or more than one source. References explicitly link to a source, like a web document containing the information on which the generated statement is based and by which it is grounded. The evaluation commences ad hoc, i.e., with a single-query and without session-based or conversational elements.

Evaluation Paradigms. To estimate the effectiveness of retrieval systems, offline evaluation within a Cranfield-style evaluation setting [22] is the de facto approach in IR research. It attempts to estimate the satisfaction of users with the output of a system by relying on an initial pool of documents judged by assessors for a given topic set [114]. These initial annotations are then be reused in throughout experiments by matching document and query identifiers. This form of evaluation offers a way to rapidly and cheaply perform large-scale evaluations of search systems. Yet, the output that generative systems produce is novel at query time. In turn, this renders it difficult to measure using such offline test collections, since no stable document identifiers are available. At its core, this is similar to the unjudged document problem. Traditionally, it is solved by assuming non-relevance [39], which is not feasible for generative IR: since all text is potentially novel, systems would not be separable through assuming non-relevance alone. Therefore, more sophisticated transfer methods are required to adapt offline evaluation for generative retrieval.

Alternatively, evaluation of generative systems can be conducted in an online fashion [110]. Here, for each run, i.e., system configuration, all output is judged anew, without relying on previous data. Yet, the effort required to judge runs during structured experimentation is immense. It requires collecting explicit user feedback about a system [58], e.g., by rating their satisfaction. Yet, it is often uncontrolled, expensive to conduct, requires time to undertake and is challenging to replicate, repeat, and reproduce [104]. Especially in an academic setting, where access to human user data is limited, much research went into simulated agents to analyze (interactive) information systems [13, 82, 83]. However, these cannot compete with "real" human feedback, which remains challenging and expensive to collect. Automatic evaluation, where the output of one model is judged by another, has been proposed as a possible way forward [74, 140], but judging the output of generative models by means of other models has itself been criticized [6, 35, 108].

3.2 Components of the User Search Process

To derive suitable evaluation objectives, first, we have to consider the search process a user undergoes when performing an ad hoc search task. Specifically, the synthetical search task enabled by generative systems should be reflected here. Based on Vakkari [128], a users' process encompasses four steps: search formulation, source selection, source interaction, and synthesis of information. Each of these can be mapped to capabilities of generative IR systems.

First, during Formulation, the user crafts a specific query that expresses the desired search outcome, addressing their information need. This is no different in generative IR systems, though what information retrieval calls a 'query' is called a 'prompt' in artificial intelligence research. To avoid confusion, we stick to the term 'query'. For the purposes of this paper, we leave this step entirely to the user who (iteratively) adapts their search formulation. Yet,

we do acknowledge that this task 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 & query expansion in traditional retrieval.

Second, during Selection, the user is presented with a result list and can then examine each entry, possibly through surrogates like snippets. The user can assess whether the results presented by the system and their information need align and thus build a focused selection of sources. In generative IR systems, this stage corresponds to the system selecting sources that contain potentially relevant information.

Third, during Interaction, the user analyzes the content of each previously selected result in-depth. The aim is to extract and structure the relevant information from each source that addresses the knowledge gap that their information need stems from. In generative IR systems, this step is supported by the model attending to relevant pieces of information previously retrieved.

Finally, during Synthesis, the user assembles the search outcome. They combine relevant information identified in multiple sources into a coherent answer to their query. In generative IR, this corresponds to the inference of the response text addressing all aspects of the user's query with information from the previously selected sources. This is key in enabling the synthetical search task. Note that interaction and synthesis often commence concurrently.

3.3 Evaluation Objectives

For each of the components of the search process, we define a corresponding evaluation objective in the context of generative IR. These are not considered evaluation steps, but rather objectives of the evaluation of the system as a whole (see Section 3.1).

Prompting Objective. Formulation is reflected in the evaluation of the models' input prompt. While search formulation is an important component to evaluate, we believe it is out of the scope of this paper, since, as previously argued, the formulation step is left to the user. For further reading, the issue of prompt engineering as an emergent field of research is covered in relevant literature on prompt engineering [42, 75, 105, 119, 122, 133, 136].

Retrieval Objective. Selection is reflected in the assessment of the context a generative IR system draws its information from. The retrieved sources (as well as any relevant information that was not retrieved) directly impact the quality of the generated response. Therefore, the retrieval objective assesses a system's ability to identify source documents satisfying a user's information need. This includes its ability to select (1) *relevant* (aligning with the users' information need), (2) *diverse* (covering a variety of information), (3) *informative* (containing valuable information), and (4) *correct* (providing accurate information) documents from a collection.

Grounding Objective. Mimicking Interaction, generative ad hoc IR models draw upon reference documents as evidence to generate a response. Yet, grounded text generation may suffer from hallucinations of broadly two types [85]: intrinsic hallucinations, where the model wrongly modifies information from the source documents, and extrinsic hallucinations, where the model generates information that is not present in the source documents. Both negatively impact the quality of the generated response [77, 85]. Therefore, the grounding objective assesses a system's ability to correlate its generated output with information from source documents. This includes its ability to (1) *identify* (find relevant information), (2) *paraphrase* (restate that information correctly), and (3) *establish consistency* (not produce contradictions to other sources).

Presentation Objective. The relevant information across multiple documents has to be synthesized into a single search outcome. This resembles multi-document summarization. Therefore, the presentation objective assesses a systems' ability to convey information to a user through the generated response in a useful manner, i.e., its ability to produce text that is (1) *concise* (at a level of granularity sensible given the topic or needed by the user [28]), (2) *coherent* (in a uniform style), and (3) *accessible* (written in an understandable way, which, again, is dependent on user needs).

3.4 Components of the User Model

Generative IR poses a challenge for developing a user model. As it is a new IR paradigm, little to no user feedback, A/B tests, laboratory studies, or user behaviour data is available in the academic context that insight about user behavior may be derived from. Further, the information search process that user behavior is traditionally grounded in is replaced (in part) by the generative system. Additionally, the assumptions of traditional user models are made for list SERPs and thus have to be revisited, taking into account the previously established evaluation objectives.

To this end, we contribute a user model for generative IR, extrapolating from established evaluation practices in related fields, like question answering, summarization, as well as traditional IR. We follow the considerations of Carterette [17], who argues that an IR-focused user model is constituted by three distinct (sub-)models: (1) a *utility* model (how each result provides utility to the user), which induces a gain function; (2) a *browsing* model (how the user interacts with results), which induces a discount function; and (3) an *accumulation* model (how the individual utility of documents is aggregated), combining the individual gain and discount values.

3.4.1 Utility Model for Generative IR. We first motivate a utility model by surveying literature on evaluation in IR and related fields. We identify 10 dimensions of utility applicable to the ad hoc synthetic search task. These are grouped into five top-level categories of *Coherence, Coverage, Consistency, Correctness* and *Clarity*. We further distinguish the unit from which gain is derived, being either an individual statement that comes from the response (statement-level) or the response as a whole (response-level). Figure 4 summarizes the dimensions of utility proposed in this section as a taxonomy, divided into response-level and statement-level dimensions; corresponding objectives are marked.

Coherence. Coherence is a response-level dimension of utility and refers to the manner in which the response is structured and presented. This includes arranging statements to form a coherent narrative without contradictions [100, 117] (*Logical Coherence*), but also a uniform style of speech (*Stylistic Coherence*), rendering it readable and engaging [16, 57]. Both implement the presentation objective at response level, amounting to "Is the response structured well?" (*Logical Coherence*) and "Does the response have a uniform style of speech?" (*Stylistic Coherence*).

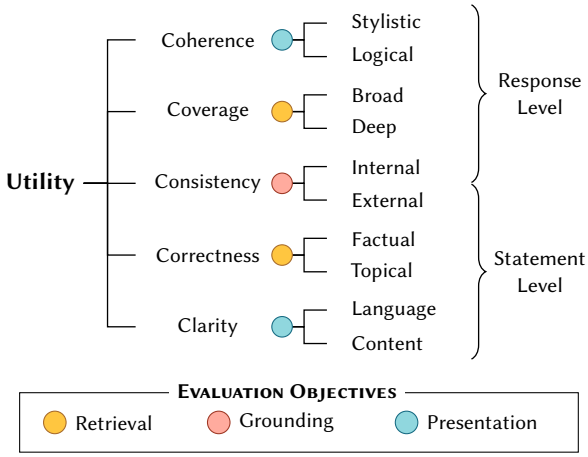

Figure 4: Taxonomy of utility dimensions in generative ad hoc retrieval; corresponding evaluation objectives colored.

Coverage. Coverage measures the cumulative extent to which presented information is pertinent to the users' information need. It can be subdivided into two forms [14]: *Broad Coverage*, i.e., whether the response covers a breadth of diverse information [148], and *Deep Coverage*, i.e., whether the response provides in-depth detailed information with high informativeness [84]. Coverage implements the retrieval objective at response level, amounting to "Does the response cover diverse information?"(*Broad Coverage*) and "Does the response offer detailed information?" (*Deep Coverage*).

Consistency. A commonly observed problem with source-based text generation is inconsistency [51] between source and generated text, which is detrimental to utility. Inconsistencies may also occur across multiple statements within a response, rendering it both a statement-level and response-level dimension. We refer to the first as *Internal Consistency* (response level), which involves assessing the consistency between statements that constitute the response, ensuring that they form a coherent answer and are not contradictory [16, 95, 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 be explained. The second, *External Consistency* (statement level), involves assessing the consistency between a statement and its source document(s), ensuring that the generated text aligns in terms of content and context [85, 108, 140]. External inconsistencies are often introduced through model hallucinations [54]. Consistency is different from factual correctness, as it only assesses the alignment of a statement with the source, and not its objective truth. Both notions implement the grounding objective but on different levels, amounting to 'Is the response free of contradictions?" (*Internal Consistency*) and "Is the statement conveying from sources accurately?" (*External Consistency*)

Correctness. Correctness gauges to which degree the information provided in the response is factually correct, 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 hard and laborious challenge [91]. It is thus often reduced to a simpler approach, framing factual correctness in terms of verifiability [74], not truth, where the main requirement is that a piece of information can attributed to a reliable reference, bestowing it correctness [37, 140]. Topical correctness denotes whether a statement aligns with the users' information need [80, 107, 137]. Both operationalize the retrieval objective at the statement level, amounting to "Does the statement state things that are verifiably true?" (*Factual Correctness*) and "Does the statement state things within the scope of the user's information need?" (*Topical Correctness*).

Clarity. The response given by a generative IR system should be expressed in a clear and understandable manner [112, 151]. This includes the use of language in a concise [28, 108], comprehensible [14] way, be lexically and grammatically correct, and accessible to the user (*Language Clarity*). 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. Orthogonal to this, the way a statement is written should always clearly communicate the most salient information [115], and where it stems from [97], in order to make the response explainable (*Content Clarity*). Both operationalize the presentation objective on the statement level, amounting to "Is a statement written in an easily readable way?" (*Language Clarity*) and "Does the statement put its focus on the most salient points?" (*Content Clarity*).

3.4.2 Reading Model for Generative IR. For list SERPs, user interaction is modeled by a browsing model, of which two fundamental kinds exist. The set-based model assumes that a user indiscriminately examines all documents given by the system, while the ranking-based model assumes a user traversing documents ascending in rank, stopping when either their information need is fulfilled or the search is aborted [17]. Aborting the search is primarily motivated by the effort being too high to justify continuing to browse. Yet, in generative IR, the selection and interaction steps of the search process are supported by the system, thus the user only has to the generated text, which requires comparably less effort. This reduces the effect of stopping criteria grounded in effort, with most users only aborting their search once their knowledge gap is fulfilled, the response is deemed insufficient, or the whole response was read. This is neither set-based, as reading the response is a sequential process and early stopping might occur, nor traditionally ranking-based, as aborting the search is not motivated by effort, but rather search satisfaction or dissatisfaction only.

We therefore propose a reading model in generative IR, as an evolution of the standard browsing model, which instead models the attention a user places on each statement while reading. Since there are no empirical studies on reading behavior for generative search at present, we instead turn to related work in reading behavior for document comprehension. We identify a total of six criteria which influence the reading process of documents for an information-seeking purpose, three of which we deem relevant to the case of generative IR. First, *Progression* [12, 69, 70, 135, 149] implies that

users parse a document sequentially, i.e., progress through the statements constituting the text in order. Second, *Decay* [38, 69, 70, 135, 149] implies that the reading attention diminishes over the span of the text. Third, *Saturation* [69, 70] implies that users abort once they read enough to fulfill their information need. In sum, this characterizes the browsing behavior of a user as sequentially reading with decaying attention, stopping early if saturated.

While three other characteristics of reading behavior have additionally been found in related work, we deem them superfluous for this reading model: (1) perceived relevance is heightened following a relevant statement [69, 149]—we adopt the restriction to a static browsing model [88, 90] without inter-statement effects, as is common ad hoc in IR evaluation. While the effect is acknowledged, its effect size may not justify its operationalization dependent on the costs; (2) attention is highest around query terms [69, 149]—we model utility not per token, but on a statement level, thus rendering this effect constant; and (3) users skip content non-relevant to them [12, 49, 69]—non-relevant statements already receive no utility.

The properties of the proposed reading model can be related to the $C/W/L$ [89] framework of browsing models for list SERPs. The conditional 'continuation probability' (C) denotes how likely a user is to continue to browse to the next item after having seen one. This can alternatively be framed in terms of the 'weight' (W), which refers to the probability of a user reaching each step of the sequence. The 'last probability' (L), indicates whether a given statement is the last one to be read before aborting, which, too, contributes to diminishing weights.

Progression indicates that the assumptions made by the $C/W/L$ framework are applicable in the first place, requiring a sequential process. Decay is encoded by a diminishing attention, relating to continuation probability and weight (C/W), while saturation relates to the last probability (L). In sum, this allows to operationalize the reading model as a monotonically decreasing weight function over statements, discounting the contribution of statements occurring later in the response. This induces a corresponding response-level document organization where the most important pieces information comes first and are then followed by increasingly insignificant details (cf. the inverted pyramid scheme of news articles [99]).

3.4.3 Accumulation Model for Generative IR. To combine gain and discount values over all considered statements, we argue in favor of the accumulation model of *expected total utility* [17, 90]. 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 [17]. However, we argue that because this effort is comparatively small in text SERPs, optimizing for it is not suitable for reliably differentiating systems in evaluation.

4 OPERATIONALIZING EVALUATION

This section considers possible 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 5, with each component (Figure rows) as a subsection in the following. The first is the experimental setup

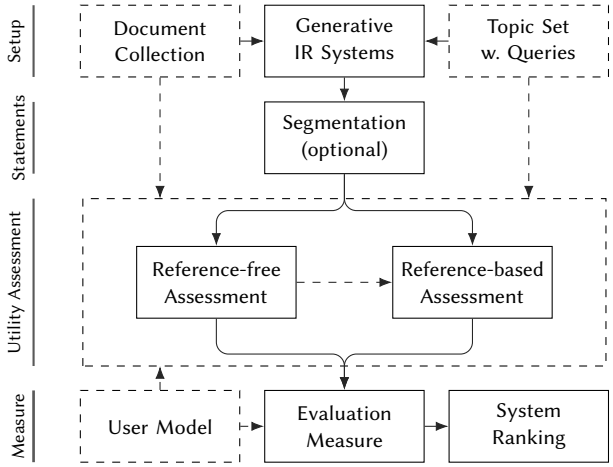

Figure 5: Overview of the evaluation procedure for generative ad hoc IR. Given documents and topics, a generative IR system produces a response, which is segmented into statements. Statements are assessed for utility in initial or repeated experimentation and an evaluation measure ranks systems by effectiveness. Solid lines indicate process flow. Dashed lines indicate contextual information sources.

(Section 4.1), encompassing a document collection, a set of topics reflecting the search task, and a set of generative IR systems to be evaluated. Their responses to queries are (optionally) split into statements using a segmentation approach (Section 4.2). Statements are then assessed for their utility (Section 4.3), distinguishing between initial experimentation without prior reference annotations, and repeated experimentation, where existing annotations can be referenced. Given annotations and an evaluation measure, the systems can then be ranked with respect to their effectiveness (Section 4.4) as indicated by an aggregated score. 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 reproducible evaluation of generative IR systems in an academic context is offline evaluation [22, 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 IR evaluation does not diverge from this basic procedure. Yet, the topics should be reflecting the actual search task generative IR systems are employed for, i.e., the synthetical task posited in Section 2.3, while ensuring that the document collection can support such queries. Furthermore, a baseline ranking of documents could be supplied for each query in order to ablate the systems' synthesizing ability, stemming from 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,[5] generative systems pose new requirements

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

here. Specifically, a run file represents text SERPs, 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 to be segmented into 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, 76, 93], and has been proposed for IR as well [108, 109]. Statements should be atomic, in the sense that an assessor should be able to make an informed and reliable decision about the utility of the statement from it and its context alone.

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 [62], 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.

4.3 Assessing Utility

Two different settings for collecting utility assessments can be discerned: (1) the response to a query is assessed solely relying on the direct assessment of the responses, without comparing to a separate ground truth; and (2) pre-existing judgments on the same document and/or query set exist to which the unjudged responses can be compared.

The first is similar to reference-free evaluation in summarization [34], which instructs annotators to assess the summary directly, while the second is similar to reference-based evaluation in summarization [9], which instructs annotators to assess the overlap between the system output and reference text, under the assumption that the reference text is the gold standard 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 to assess. For example, topical correctness requires both response and query, while factual correctness takes into account query, text, and the external sources. We therefore discern reference judgments and context: reference judgments are one or more existing assessments to which the new one is compared, while context covers the information necessary to judge. A judgment made with context only is therefore deemed reference-free. Collecting initial assessments of utility within an offline evaluation setting is laborious, since the to-be-judged texts are dynamic (text SERPs are generated at query time), and thus each new response has to be manually assessed by a judge.

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 [92] or relying on other generative models to assess the quality of the output, such as BARTScore [139] or GPTScore [40]. Classifiers trained to estimate the magnitude of a utility dimension have also been used [65]. 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. [66] evaluate their generative IR system for the task of question answering by matching words between generated response and the gold answer. Other content overlap metrics such as BLEU [98], NIST [32], ROUGE [72] TER [120], METEOR [5], BERT Score [144], or MoverScore [147] have been used to compare to a ground truth, either the full response or each statement individually. However, these measures should not be used to assess overlap with retrieved documents, as these are not an adequate ground truth source. Ranking models have also been proven useful for relative assessment of candidates to available ground-truth, e.g., in machine translation [33, 121], both in a listwise [68] as well as a pairwise setting [47, 48].

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 $\sum_{i=1}^{k} g(d_i) \cdot \sum_{j=i}^{k} p(j)$ [17], 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 prior, the latter as a discount function, chosen based on prior information about typical user behavior. The widely established measures of *DCG* and *nDCG* [53] used for traditional IR evaluation stem from this family of measures [17] and seem suitable for generative IR 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* [21] 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 IR systems have been proposed recently: SWAN [108] and EXAM [113]. This naturally yields the question how our proposed evaluation approach compares to these two. The starting point of both is a text SERP response, albeit less formalized and without considering the synthetical search task it enables.

SWAN follows a similar approach as 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 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 synthetical 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 not accepted standards.

In sum, our approach can be related to existing methods in terms of compatibility, complementarity, and consistency. It is compatible with SWAN, being derived from similar assumptions, yet adding a theoretical foundation, and constructed with a different search task in mind. It is complementary to EXAM, with a focus on fine-grained, reliable, user-oriented evaluation, whereas EXAM excels for rapid, system-oriented experimentation with little overhead. And overall, our approach is consistent with traditional IR evaluation techniques, making only small adaptations to utility, browsing, and aggregation model 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 IR systems offer a new paradigm for the retrieval of information. With this new paradigm comes the need to measure and understand the new dimensions that make text SERP responses from these systems relevant to a user's information need. In this survey, we have investigated a theoretical foundation for the evaluation of generative IR systems, extrapolated from traditional IR and related domains. Firstly, we established that the search task of generative ad hoc IR goes beyond acquiring information, and instead enables the condensation of information, a process we dub the 'synthetical search task'. The different system architectures enabling this task were shortly outlined. Given this departure from traditional ad hoc IR, we proposed a new user model that accommodates the task. Here, we also extrapolated existing frameworks to model the generative IR search process, including evaluation objectives, utility dimensions and a browsing model for text SERPs. Finally, we outlined how one could operationalize the evaluation of generative IR systems, surveying how existing evaluation approaches relate to, and could fit into the proposed methodology.

Many techniques for constructing generative IR systems are currently emerging but evaluating the output of such systems is a non-standardized and thus rarely comparable effort lacking a theoretical motivation and methodological rigor. We have provided in this paper our vision of a comprehensive approach for evaluating generative ad hoc IR systems. We firmly believe that this survey provides the IR community with the foundation to conduct future research into new methods for the evaluation of generative ad hoc IR. Yet, we also have several directions of future work planned, and several open questions to tackle. Near-future work includes conducting a rigorous empirical evaluation based on our proposal, and studying its reliability and validity within user studies. We believe that user experiments are required to effectively apply the theoretical motivation developed in this survey. We plan a meta-evaluation of both existing measures and measures modified for generative IR specifically, to study how well they align with user preferences. We also plan to study the proposed utility dimensions and their ability to reflect user satisfaction, akin to studies conducted for traditional IR [14]. In addition, investigating the way users interact with generative retrieval systems is warranted; for example, do clicks indicate relevance as before, or rather the opposite, with the aim of generative ad hoc IR being to 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 IR system individually, we focus on evaluating the final response. Second, the evaluation is additionally limited to answer the question if an generative ad hoc IR system is successful at supporting the synthetical 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 ad hoc IR, while precluding 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 received funding from the European Union's Horizon Europe research and innovation programme under grant agreement № 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014). The authors also acknowledge financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project ID ScaDS.AI. Harrisen Scells is the recipient of an Alexander von Humboldt Stiftung Research Fellowship.

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. https://doi.org/10.4230/DagRep.3.10.92

[2] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[3] Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. 2023. GAR-meets-RAG Paradigm for Zero-Shot Information Retrieval. *CoRR* abs/2310.20158 (2023). https://doi.org/10.48550/ARXIV.2310.20158 arXiv:2310.20158

[4] Lorena Leal Bando, Falk Scholer, and Andrew Turpin. 2010. Constructing query-biased summaries: a comparison of human and system generated snippets. In *Information Interaction in Context Symposium, IIiX 2010, New Brunswick, NJ, USA, August 18-21, 2010*, Nicholas J. Belkin and Diane Kelly (Eds.). ACM, 195–204. https://doi.org/10.1145/1840784.1840813

[5] 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. https://aclanthology.org/W05-0909/

[6] Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr. 2023. Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. *CoRR* abs/2305.01509 (2023). https://doi.org/10.48550/arXiv.2305.01509 arXiv:2305.01509

[7] Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-IR @ SIGIR 2023: The First Workshop on Generative Information Retrieval. *CoRR* abs/2306.02887 (2023). https://doi.org/10.48550/arXiv.2306.02887 arXiv:2306.02887

[8] 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 *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/cd88d62a2063fdaf7ce6f9068fb15dcd-Abstract-Conference.html

[9] 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. https://doi.org/10.18653/v1/2020.emnlp-main.751

[10] 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. https://proceedings.mlr.press/v162/borgeaud22a.html

[11] Andrei Z. Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (2002), 3–10. https://doi.org/10.1145/792550.792552

[12] 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. https://doi.org/10.1145/2070719.2070722

[13] Arthur Câmara, David Maxwell, and Claudia Hauff. 2022. Searching, Learning, and Subtopic Ordering: A Simulation-based Analysis. (2022). https://dblp.org/rec/journals/corr/abs-2201-11181

[14] 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. https://doi.org/10.1145/3406522.3446028

[15] 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. https://openreview.net/forum?id=5k8F6UU39V

[16] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? *CoRR* abs/2307.03826 (2023). https://doi.org/10.48550/arXiv.2307.03826 arXiv:2307.03826

[17] 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. https://doi.org/10.1145/2009916.2010037

[18] Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'Grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4283–4305. https://doi.org/10.18653/v1/2021.findings-acl.375

[19] 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. https://doi.org/10.1145/3477495.3531827

[20] Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Abstractive Snippet Generation. In *Web Conference (WWW 2020)*, Yennung Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM, 1309–1319. https://doi.org/10.1145/3366423.3380206

[21] 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. https://doi.org/10.1145/1390334.1390446

[22] Cyril W. Cleverdon. 1997. The Cranfield tests on index language devices.

[23] The AutoGPT Contributors. 2023. AutoGPT: The Heart of the Open-Source Agent Ecosystem. https://github.com/Significant-Gravitas/AutoGPT.

[24] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[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). https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf

[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). https://trec.nist.gov/pubs/trec31/papers/Overview_deep.pdf

[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. https://doi.org/10.1145/3274784.3274788

[28] Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the document understanding conference*, Vol. 2005. 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. https://aclanthology.org/P07-1097/

[30] Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 Question Answering Track 99. 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). http://trec.nist.gov/pubs/trec15/papers/QA06.OVERVIEW.pdf

[31] 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 *CHIIR*. ACM, 172–186.

[32] George R. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.

[33] Kevin Duh. 2008. Ranking vs. Regression in Machine Translation Evaluation. In *WMT@ACL*.

[34] 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. https://doi.org/10.1162/tacl_a_00373

[35] 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. https://doi.org/10.1145/3578337.3605136

[36] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086* (2021).

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

[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), 39.

[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. https://doi.org/10.1007/978-3-031-28244-7_20

[40] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. *CoRR* abs/2302.04166 (2023). https://doi.org/10.48550/arXiv.2302.04166 arXiv:2302.04166

[41] Luke Gallagher, Marwah Alaofi, Mark Sanderson, and Falk Scholer. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. https://www.microsoft.com/en-us/research/uploads/prod/2023/05/srp0313-alaofi.pdf

[42] 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. https://doi.org/10.18653/v1/2021.acl-long.295

[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. https://doi.org/10.18653/v1/2020.acl-main.511

[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. https://doi.org/10.1007/978-3-031-28238-6_31

[45] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *CoRR* abs/2209.12356 (2022). https://doi.org/10.48550/arXiv.2209.12356 arXiv:2209.12356

[46] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *CoRR* abs/2002.08909 (2020). arXiv:2002.08909 https://arxiv.org/abs/2002.08909

[47] Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez i Villodre, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to Differentiate Better from Worse Translations. In *Conference on Empirical Methods in Natural Language Processing*.

[48] Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez i Villodre, and Preslav Nakov. 2015. Pairwise Neural Machine Translation Evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

[49] 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. https://doi.org/10.1145/2637002.2637011

[50] Marti A. Hearst. 2009. *Search User Interfaces*. Cambridge University Press. https://doi.org/10.1017/CBO9781139644082

[51] Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. 2021. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. *CoRR* abs/2104.14839 (2021). arXiv:2104.14839 https://arxiv.org/abs/2104.14839

[52] 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. https://doi.org/10.18653/v1/2021.eacl-main.74

[53] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[54] 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. https://doi.org/10.1145/3571730

[55] Zhengbao Jiang, Luyu Gao, Jun Araki, Haibo Ding, Zhiruo Wang, Jamie Callan, and Graham Neubig. 2022. Retrieval as Attention: End-to-end Learning of Retrieval and Reading within a Single Transformer. (2022). https://dblp.org/rec/journals/corr/abs-2212-02027

[56] 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. *CoRR* abs/2305.06983 (2023). https://doi.org/10.48550/arXiv.2305.06983 arXiv:2305.06983

[57] 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. https://doi.org/10.18653/v1/2020.acl-main.456

[58] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[59] 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. (2022). https://dblp.org/rec/journals/corr/abs-2212-14024

[60] 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. https://openreview.net/pdf?id=_nGgzQjzaRy

[61] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *41st International ACM Conference on Research and Development in Information Retrieval (SIGIR 2018)*. ACM, 1257–1260. https://doi.org/10.1145/3209978.3210160

[62] 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. https://doi.org/10.1007/978-3-030-72240-1_5

[63] Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. 2021. Meta-Information in Conversational Search. *ACM Transactions on Information Systems (ACM TOIS)* 39, 4, Article 50 (Aug. 2021), 44 pages. https://doi.org/10.1145/3468868

[64] 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), 1–16.

[65] Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

[66] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *CoRR* abs/2203.05115 (2022). https://doi.org/10.48550/arXiv.2203.05115 arXiv:2203.05115

[67] 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.). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[68] Maoxi Li, Aiwen Jiang, and Mingwen Wang. 2013. Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation. In *Machine Translation Summit*.

[69] 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. https://doi.org/10.1145/3269206.3271764

[70] 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. https://doi.org/10.1145/3331184.3331205

[71] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. (2008), 1822–1837. https://dblp.org/rec/journals/ipm/LiB08

[72] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[73] 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. https://doi.org/10.1561/1500000073

[74] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. *CoRR* abs/2304.09848 (2023). https://doi.org/10.48550/arXiv.2304.09848 arXiv:2304.09848

[75] 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. https://doi.org/10.1145/3491102.3501825

[76] 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. https://doi.org/10.18653/v1/2023.acl-long.228

[77] Klaus-Michael Lux, Maya Sappelli, and Martha 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*. Association for Computational Linguistics, Online, 1–10. https://doi.org/10.18653/v1/2020.eval4nlp-1.1

[78] Sean MacAvaney, Craig Macdonald, Roderick Murray-Smith, and Iadh Ounis. 2021. IntenT5: Search Result Diversification using Causal Language Models. *CoRR* abs/2108.04026 (2021). arXiv:2108.04026 https://arxiv.org/abs/2108.04026

[79] 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*. 1573–1576.

[80] 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. https://doi.org/10.1145/3002172

[81] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge.

[82] David Maxwell and Leif Azzopardi. 2016. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour.. In *CIKM*. 731–740. https://dblp.org/rec/conf/cikm/MaxwellA16

[83] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies.. In *CIKM*. 313–322. https://dblp.org/rec/conf/cikm/MaxwellAJK15

[84] 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. https://doi.org/10.1145/3077136.3080824

[85] 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. https://doi.org/10.18653/v1/2020.acl-main.173

[86] 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. *CoRR* abs/2302.07842 (2023). https://doi.org/10.48550/arXiv.2302.07842 arXiv:2302.07842

[87] 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.

[88] 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. https://doi.org/10.1145/2838931.2838938

[89] 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. https://doi.org/10.1145/3052768

[90] 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. https://doi.org/10.1145/2505515.2507665

[91] 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. https://doi.org/10.1007/978-3-030-85251-1_19

[92] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. 2019. Do Deep Generative Models Know What They Don't Know?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1xwNhCcYm

[93] 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. https://doi.org/10.1145/1233912.1233913

[94] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[95] 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. https://doi.org/10.18653/v1/D19-1315

[96] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019). arXiv:1904.08375 http://arxiv.org/abs/1904.08375

[97] 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. https://ojs.aaai.org/index.php/HCOMP/article/view/5284

[98] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

[99] Horst Pöttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies* 4, 4 (2003), 501–511. https://doi.org/10.1080/1461670032000136596

[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 Human Information Interaction and Retrieval*. 117–126. https://doi.org/10.1145/3020165.3020183

[102] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *CoRR* abs/2302.00083 (2023). https://doi.org/10.48550/arXiv.2302.00083 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. https://doi.org/10.1145/3539618.3591890

[104] Gareth Renaud and Leif Azzopardi. 2012. SCAMP: a tool for conducting interactive information retrieval experiments.. In *IIiX*. 286–289. https://dblp.org/rec/conf/iiix/RenaudA12

[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. https://doi.org/10.1145/3411763.3451760

[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. https://openreview.net/pdf?id=yKbprarjc5B

[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. https://doi.org/10.1145/3209978.3210052

[108] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. *CoRR* abs/2305.08290 (2023). https://doi.org/10.48550/arXiv.2305.08290 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. https://doi.org/10.1145/2063576.2063669

[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).

[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. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2016).

[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. https://ceur-ws.org/Vol-2950/paper-16.pdf

[114] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 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. https://doi.org/10.1145/3531146.3533127

[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 preprint arXiv:2305.14292* (2023).

[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. https://doi.org/10.18653/v1/2021.naacl-main.411

[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. *CoRR* abs/2301.12652 (2023). https://doi.org/10.48550/arXiv.2301.12652 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. https://doi.org/10.18653/v1/2020.emnlp-main.346

[120] Matthew G. Snover, B. Dorr, R. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Conference of the Association for Machine Translation in the Americas*.

[121] Xingyi Song and Trevor Cohn. 2011. Regression and Ranking based Optimisation for Sentence Level MT Evaluation. In *WMT@EMNLP*.

[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. https://doi.org/10.18653/v1/2022.acl-long.60

[123] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *CoRR* abs/2304.09542 (2023). https://doi.org/10.48550/arXiv.2304.09542 arXiv:2304.09542

[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 *NeurIPS*.

[125] James Thorne. 2022. Data-Efficient Auto-Regressive Document Retrieval for Fact Verification. *ArXiv* abs/2211.09388 (2022).

[126] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 2–10. https://doi.org/10.1145/290941.290947

[127] Johanne R. Trippas, Damiano Spina, Paul Thomas, Hideo Joho, Mark Sanderson, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2020), 1–19. https://doi.org/10.1016/j.ipm.2019.102162

[128] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *J. Inf. Sci.* 42, 1 (2016), 7–18. https://doi.org/10.1177/0165551515615833

[129] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2020. An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR 2020)*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2085–2088. https://doi.org/10.1145/3397271.3401297

[130] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *European Conference on Information Retrieval*. Springer, 541–557. https://doi.org/10.1007/978-3-030-15712-8_35

[131] 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. https://doi.org/10.18653/v1/2020.emnlp-main.609

[132] 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 *NeurIPS*.

[133] 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. *CoRR* abs/2302.11382 (2023). https://doi.org/10.48550/arXiv.2302.11382 arXiv:2302.11382

[134] 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.

[135] 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. https://doi.org/10.1145/3543507.3583343

[136] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. *CoRR* abs/2309.03409 (2023). https://doi.org/10.48550/arXiv.2309.03409 arXiv:2309.03409

[137] 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. https://doi.org/10.1145/3077136.3084154

[138] 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. https://openreview.net/pdf?id=fB0hRu9GZUS

[139] 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. https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html

[140] Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. *CoRR* abs/2305.06311 (2023). https://doi.org/10.48550/arXiv.2305.06311 arXiv:2305.06311

[141] 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)*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1181–1190. https://doi.org/10.1145/3397271.3401160

[142] Dake Zhang and Ronak Pradeep. 2023. ReadProbe: A Demo of Retrieval-Enhanced Large Language Models to Support Lateral Reading. *CoRR* abs/2306.07875 (2023). https://doi.org/10.48550/arXiv.2306.07875 arXiv:2306.07875

[143] 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 *44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2521–2525.

[144] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*. OpenReview.net.

[145] 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. https://doi.org/10.1145/3437963.3441758

[146] 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. *CoRR* abs/2304.11076 (2023). https://doi.org/10.48550/arXiv.2304.11076 arXiv:2304.11076

[147] 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. https://doi.org/10.18653/v1/D19-1053

[148] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. 2012. Coverage-based search result diversification. *Inf. Retr.* 15, 5 (2012), 433–457. https://doi.org/10.1007/s10791-011-9178-4

[149] 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. https://doi.org/10.1145/3331184.3331231

[150] Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, and Ji-Rong Wen. 2022. DynamicRetriever: A Pre-training Model-based IR System with Neither Sparse nor Dense Index. *CoRR* abs/2203.00537 (2022).

[151] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2009. A multi-dimensional model for assessing the quality of answers in social Q&A sites. In *ICIQ*. 264–265.

[152] 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. *CoRR* abs/2206.10128 (2022).

[153] Shengyao Zhuang and Guido Zuccon. 2021. Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. *arXiv preprint arXiv:2108.08513* (2021).

[154] Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large Language Models are Built-in Autoregressive Search Engines. arXiv:2305.09612 [cs.CL] http://arxiv.org/abs/2305.09612v1

[155] Guido Zuccon and Bevan Koopman. 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793* (2023).