On the Evaluation of Machine-Generated Reports

James Mayfield
Johns Hopkins University
Baltimore, MD, USA
mayfield@jhu.edu

Eugene Yang
Johns Hopkins University
Baltimore, MD, USA
eugene.yang@jhu.edu

Dawn Lawrie
Johns Hopkins University
Baltimore, MD, USA
lawrie@jhu.edu

Sean MacAvaney
University of Glasgow
Glasgow, United Kingdom
sean.macavaney@glasgow.ac.uk

Paul McNamee
Johns Hopkins University
Baltimore, MD, USA
mcnamee@jhu.edu

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

Luca Soldaini
Allen Institute for AI
Seattle, WA, USA
luca@soldaini.net

Ian Soboroff
NIST
Gaithersburg, MD, USA
ian.soboroff@nist.gov

Orion Weller
Johns Hopkins University
Baltimore, MD, USA
oweller2@jhu.edu

Efsun Kayi
Johns Hopkins University
Baltimore, MD, USA
ekayi1@jhu.edu

Kate Sanders
Johns Hopkins University
Baltimore, MD, USA
ksande25@jhu.edu

Marc Mason
Johns Hopkins University
Baltimore, MD, USA
mmason8@jhu.edu

Noah Hibbler
University of Maryland
College Park, MD, USA
ndhibbl@terpmail.umd.edu

ABSTRACT

Large Language Models (LLMs) have enabled new ways to satisfy information needs. Although great strides have been made in applying them to settings like document ranking and short-form text generation, they still struggle to compose complete, accurate, and verifiable long-form reports. Reports with these qualities are necessary to satisfy the complex, nuanced, or multi-faceted information needs of users. In this perspective paper, we draw together opinions from industry and academia, and from a variety of related research areas, to present our vision for automatic report generation, and—critically—a flexible framework by which such reports can be evaluated. In contrast with other summarization tasks, automatic report generation starts with a detailed description of an information need, stating the necessary background, requirements, and scope of the report. Further, the generated reports should be complete, accurate, and verifiable. These qualities, which are desirable—if not required—in many analytic report-writing settings, require rethinking how to build and evaluate systems that exhibit these qualities. To foster new efforts in building these systems, we present an evaluation framework that draws on ideas found in various evaluations. To test completeness and accuracy, the framework uses nuggets of information, expressed as questions and answers, that need to be part of any high-quality generated report. Additionally, evaluation of citations that map claims made in the report to their source documents ensures verifiability.

SIGIR '24, July 14–18, 2024, Washington, DC, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657846

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Evaluation, Report Generation, Text Analysis, Factual Citation

ACM Reference Format:
James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the Evaluation of Machine-Generated Reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3626772.3657846

1 INTRODUCTION

The emergence of generative Large Language Models (LLMs) has brought with it the ability to automatically generate all kinds of text. With it, a host of problems—old and new—have (re)emerged that affect these generated texts. The fields of Information Retrieval (IR) and Natural Language Processing (NLP) both have important roles in building new methods to improve text generation and in designing approaches to evaluate the quality of these methods.

LLMs can enable new ways for people to satisfy various information needs. Simple information needs (e.g., factoids) can be answered with relatively short generated responses pointing to a single source. However, when information needs are complex, nuanced, or multifaceted, a suitable response must also be more complex. They need to draw together numerous facts gathered from potentially multiple sources to completely and faithfully respond to the information need. We refer to this longer-form answer generation as a "report" on a user-specified topic.

More formally, we define a report as a text that attempts to satisfy an explicitly stated information need by finding documents in a corpus (potentially a mixture of text, images, tables, etc.) that contain relevant information, expressing that information in the text, and providing appropriate citations from the report to the supporting documents. We envision a high-quality report as the *ideal* response to a user with a complex task in mind, since such a report would succinctly, coherently, and verifiably cover all the information in a corpus pertinent to their information need. Note that this definition makes the framework better suited to reports that inform an analyst than to reports that generate novel analyses.

Report writing can be viewed as a natural downstream task of Retrieval Augmented Generation (RAG), where faithfulness has been a focus of study [52, 73, 76]. In this view, an LLM generates the report using the report request as part of the prompt and searches the document collection for relevant information that can be added to the prompt to ensure the report's accuracy. Report generation can also be thought of as summarization. From the summarization viewpoint, a report is an attributed task-based informative abstractive multi-document summary (see Section 3.1.1 for a more detailed explanation of these categories). Such a report might also include portions that are not summaries at all, but are, for example, introductory material or comparisons of the summarized information.

We posit that all of these viewpoints are valid, and each informs evaluation for report generation. This work describes an abstract framework for evaluating automated report generation, ARGUE (Automated Report Generation Under Evaluation), that is built on top of lessons learned from prior evaluation approaches in information retrieval, summarization and text generation. It will be used by the TREC track NeuCLIR in its report generation task.[1] The ARGUE framework builds a foundation for a broader research agenda in evaluating automatically generated long-form text beyond reports.

Some of ARGUE's most important features are:

- We use the concept of information **nuggets** out of the summarization literature to capture the content a report should contain. We express each nugget as a question together with a list of acceptable answers to that question.
- **Citations** are a key report component. A citation is a pointer from a source element in the report (typically a sentence) to a target element in a document (typically the entire document).
- We propose that precision and recall serve as the basis for most content-based measures. ARGUE supports precision measures over the sentences of the report, and recall measures over the information nuggets.

2 REQUIREMENTS

This section defines requirements of a report evaluation system.

We first define the various actors (and one non-actor) in ARGUE:

Report Requester: The person requesting the report. This is the person whose purpose the report should satisfy.

Report Audience: The person who will be reading the report. This is often the same as the report requester.

Report Writer: The automated system that takes a report request and a document collection as inputs and produces the report.

Report Request: A detailed specification of the report to be written. The report request can include:

- *User story:* explains the report requester's background, situation, and report-writing philosophy, as well as a description of the audience for the report.
- *Problem statement:* indicates the content that the report is required to contain.
- *Background*: describes what is already known about the topic that need not appear in the report.
- *Constraints:* specifies restrictions such as the length of the report or a temporal window for sources.

Assessor: Any person making judgments in producing evaluation materials or scoring submitted runs. Assessors include those selecting report topics, writing report requests, identifying nuggets, binding nuggets to documents in the collection, and making other judgments necessary to assign scores to reports.

The evaluation we advocate has several key attributes. First, it must ensure that the report is responsive to the report request. It must ensure the report's key information presented is attested in the document collection, that the report properly cites those documents, and that the information they contain is faithfully captured by the report. It must score a report using evaluation data created by a person. While scoring may be automated, requiring the ground truth data to be human-generated helps to prevent circularity between report generation and report evaluation, thereby reducing the bias the evaluation might have toward e.g., a particular generative model. Finally, the evaluation must have the intention of reusability. Producing a reusable evaluation is challenging because of the level of interpretation required to make the required judgments. Reusability is thus often at odds with the other goals of an evaluation. The information retrieval community has thought through many of the issues underlying reusability, and we present ARGUE to try to take advantage of that experience.

While it is nearly impossible to accurately claim that any evaluation component is novel, there are points of emphasis in our proposed evaluation style that we think make it stand out from other extant text generation evaluations. First is the amount and detail of the background information provided in the report request. While other evaluations have provided additional information describing inclusion criteria, in practice systems have often focused only on brief specifications. For example, a narrative giving detailed information about what should and should not be considered relevant, long a part of TREC topics, has rarely been exploited. The arrival of large language models that can easily incorporate such materials makes now an opportune time to focus on including ancillary documentation in a report request, not just for this style of evaluation, but for any text generation evaluation. While we advocate that these ancillary details be made explicit in the evaluation, we acknowledge that in real report-writing applications implicit knowledge might be more practical and adequate for the task.

Second, until recently hallucination in text generation system output was not a major focus, primarily because generative systems were not good enough to create convincing hallucinated text. With the rise of large generative LLMs hallucination has become a common part of text generation system output; the evaluation must account for this as well.

Borrowing from an IR evaluation perspective, we promote the view of nuggets as opinion, not fact. In report evaluation, nuggets play the role that relevant documents play in IR. Were document

[1] https://neuclir.github.io/

relevance treated as fact rather than opinion, it would be virtually impossible to come to agreement on which documents were relevant to a given topic; inter-annotator agreement would be too low. Treating relevance as opinion avoids this problem. In exchange, relevance as opinion adds constraints to the evaluation, primarily that the author of the topic should be the relevance assessor. If relevance is not decided until after system submissions, that means that assessor continuity is important; assessors should be selected such that they can create topics at one time, and assess relevance at a later time, possibly months later. We advocate accepting this tradeoff for nuggets in report generation evaluation. For nuggets, the implication is that items reasonably seen by a report writer as nuggets might not be identified in advance by the assessor. A given evaluation might address this issue through a pyramid approach [61] to identify nugget importance if multiple reference reports are available. Or an evaluation might determine that nugget subjectivity will not change the preference order of meaningfully different systems and ignore it. In either case, we recommend that report sentences bearing and accurately reflecting a citation should not be penalized during scoring, precisely because they might be valid nuggets in someone's eyes. Constraints such as maximum document length can discourage intentional overgeneration of sentences that have a small chance of matching assessor nuggets.

To meet these requirements, four broad questions should be asked about each report being evaluated:

Q1 Does the report include the information contained in the document collection that the report requires?
Q2 Does it accurately express all such information?
Q3 Does it contain appropriate citations to the collection?
Q4 Has the information been fitted together into a useful form?

Q4 is a crucial part of any text generation evaluation. It covers such attributes as fluency [65], coherence [40, 50], consistency [32], and rhetorical structure [16, 21]. In light of this importance, it has a long history and has been studied in depth elsewhere. Thus, while we leave a place for this in the overall evaluation in ARGUE, we leave it to others to address it in light of the changing NLP landscape.

3 BACKGROUND

Here we review related work on report writing and evaluation.

3.1 Report Writing

Report writing involves text generation, for which prior work on summarization and RAG provides useful perspectives.

3.1.1 *Summarization.*

In its most general form, a summary is a document whose substantive content is based entirely on the content of other *target document(s),* and that is more concise than simply presenting the other document(s) in their original form would have been [55]. Summaries have been defined along several axes:

- Single-document or Multi-document [47]: Is the summary built from one document (single-document), or many (multi-document)?
- Extractive or Abstractive [13]: Does the summary primarily draw language from the summarized documents (extractive), or does it generate new language (abstractive)?
- Indicative or Informative [37]: Does the summary help the reader to decide whether to read the summarized document(s) (indicative), or does it include enough content to make it unnecessary to read those document(s) (informative)?
- Generic or Task-Based [83]: Is the summary constructed with no particular task in mind (generic), or is there a specific task that the summary is designed to support (task-based)?
- Attributed or Unattributed [70]: Does the summary include citations to the summarized documents (attributed), or does it lack citations (unattributed)?
- Original or Update [56, 63]: Should the summary include all information (original), or only information that the reader does not already know (update)?
- Closed or Open Domain [26, 92]: Are the documents to summarize supplied (closed domain), or must the system perform a search to identify the appropriate documents (open domain)?

The reports in which we are interested are attributed task-based informative abstractive open-domain multi-document summaries that may call for either original or update summaries.

3.1.2 *Retrieval-Augmented Generation.*

Following preliminary research on furnishing transformer architectures with external knowledge sources, Lewis et al. [41] introduce RAG models as a way to improve language model performance on knowledge-intensive tasks, using an encoded Wikipedia collection as a non-parametric memory system. RAG models have since been used to improve dialogue systems [38, 77], machine translation [7, 8], and text-style transfer [44] among other applications [43].

Various approaches have been proposed to incorporate RAG models into summarization [2, 64] and other document generation tasks. One use of retrieval has been to find an example summary, sometimes with retrieved summary reranking [9], to serve as a template for the summary of another document. Retrieval can also be used to improve language model factuality. By curating large, high quality collections, generation can be grounded in supporting documents [4]. This mechanism has been shown to be particularly beneficial for rarer entities and concepts [54]. Finally, RAG enables LLMs to access information that was not available at pre-training time, such as proprietary or copyrighted information [57].

Vision-language modeling [1, 3, 51] enables multimodal retrieval-augmented generation systems that benefit from rich non-textual data [33, 60]. Different modalities facilitate the completion of different tasks, including image understanding [12, 93], open-domain VQA [33, 49], translation [20], and multimodal generation [89].

3.2 Evaluation

As report generation includes elements of several prior tasks, including document retrieval, summarization, question answering, and retrieval-augmented generation, we briefly review salient work on those tasks that we see as related to ARGUE.

3.2.1 *Information Retrieval.*

Evaluation of ad hoc retrieval is typically based on assessor-produced relevance judgments of documents that are selected by pooling system responses in a shared task, or sometimes based on active learning [29, 72]. Obtaining both good precision and good recall is important in real-world systems, so commonly used metrics combine both components (e.g., mean average precision, nDCG [35]). Statistical significance testing can be performed, for example with Student's *t*-test [78].

In a report-writing scenario, recall is important to allow assessment of how comprehensively the report responds to the report request. Precision is also important for automated report generation; reports are a type of multi-document synthesis, and incorporating content from non-pertinent documents can adversely affect the utility of the report.

To create evaluation datasets for report writing, care must be taken to develop report requests that match information available in the document collection. If requests are too broadly scoped, or if too much salient information is present in the collection, it will be difficult (i.e., prohibitively expensive in human labor) to determine the full set of correct nuggets present in the collection.

3.2.2 Summarization. Evaluating automatic summarization can require significant manual effort. In 2001, NIST initiated the Document Understanding Conference (DUC) to develop evaluation methods for summarization. DUC continued until 2007 and then became the summarization track of the Text Analysis Conference (TAC) through 2014. The DUC/TAC summarization evaluations were notable for having people write summaries manually, and using those "model" summaries (or "reference texts") as the jumping-off point for metric development.

The DUC evaluation procedure measured coverage (that is, recall) through a pairwise comparison between two summaries: the model summary and a "peer" summary (which could be a generated summary or another model). The model was divided into *Elementary Discourse Units* (EDUs), essentially clauses [45, 80] while the peer was split on sentence boundaries. An assessor would match each EDU with the sentences in the peer that contained that information, and indicate how much of the meaning of the EDU was expressed in the corresponding matched peer units. Unmarked sentences in the peer were then marked for relevance. Harman and Over [31] found that model summaries from different authors were markedly different, and that assessors also did not agree on model unit coverage ratings.

Work also began around DUC 2003 on automatic metrics, specifically comparing the model summary to the peer using word n-gram statistics. Lin and Hovy [48] looked at the BLEU measure developed for machine translation, and found that recall on word unigrams correlated better with the DUC assessments than full BLEU scoring, which incorporates longer n-grams. Following that, they developed ROUGE [46], a recall-oriented metric similar to BLEU. ROUGE has a number of variants depending on how tokens are parsed, how n-grams are selected and assembled, and how scores are aggregated across summaries to obtain a system score. A study by Graham [28] explored a large grid of ROUGE parameters in comparison with BLEU using data from DUC-2004, and found that BLEU and ROUGE-2 (2-grams, stemmed, stopwords removed, computing an average of precision scores) had the highest correlation with human assessment. ROUGE has been used to evaluate summarization [46], Long-Form Question Answering (LFQA) [39, 88] and RAG [41]. ROUGE has well-documented problems as an evaluation metric in e.g., summarization [28] or LFQA [39]. From our perspective, its main problems as an evaluation metric for report generation are its requirement for reference reports (making it expensive), its poor robustness to hallucination (making it inaccurate), and that it does not handle citations (making it incomplete).

In 2004, Nenkova and Passonneau [61] proposed the "Pyramid Method" for evaluation. Since comparing generated summaries against a model is subject to the inherent variation in model summaries, they propose to abstract the model summaries into *Summary Content Units* (SCUs). SCUs are clauses that appear (with more or less the same meaning) in multiple model summaries. They are weighted by the number of model summaries that express them. Figure 1 shows an example of two SCUs from parts of four model summaries.

In informal usage, SCUs have been referred to as "nuggets." Rather than being a clause, a nugget might be a description of a concept along with how it was expressed in the models.[2] Subsequent research on the pyramid method has focused on automatic creation and alignment of SCUs. For example, Gao et al. [24] performs a dependency parse of the model summary, then represents individual clauses using vector embeddings. Nugget fuzziness can be addressed by using hoppers [59, 79] to bin together differing descriptions that refer to the same item.

The main difficulties in using nuggets for report evaluation are that they treat hallucinations (contradictions and misinformation) exactly the same as content that has no matching nugget, and that they do not support citations. We have incorporated nugget-based evaluation into ARGUE, tying nuggets to reports not directly, but rather through cited documents.

3.2.3 Question Answering. Factoid Question Answering (QA) evaluation typically consists of using accuracy or F_1 against a gold standard answer (or answer set) [15, 69, 84]. This type of evaluation has many advantages, as it can be easily automated and is simple to annotate. Long-form QA [19, 62] is evaluated similarly to summarization, typically with automated metrics like ROUGE, model-based metrics like BERTScore [91] or BLEURT [75], or human evaluation [39, 88].

3.2.4 Retrieval-Augmented Generation. Early retrieval augmented generation systems have been evaluated using task-specific metrics on end-to-end tasks. For example, in the context of question answering, exact match and F_1 metrics have been used [30, 41]. For summarization, ROUGE and BERTScore on reference summaries are common [26]. These approaches have two limitations: they only measure ability to complete end tasks, and thus cannot assess intermediate stages or evaluate generation across multiple dimensions; and they are not well-suited to capture failures that can be introduced by current generative models [27].

More recently, techniques have proposed to more holistically evaluate RAG systems. Gienapp et al. [25] introduce a theoretical framework for evaluating ad hoc generative retrieval. Chen et al. [11] focus on robustness of RAG systems against various perturbations. Thakur et al. [82] benchmark hallucinations and the ability of RAG systems to identify relevant information for 18 languages. Others have introduced benchmarks to measure the ability of RAG systems to provide citations [6, 23, 53, 90]. While not specifically

[2] See https://tac.nist.gov/publications/2010/presentations/TAC2010_Summ_Overview.pdf for an example of SCUs as nuggets.

A1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya. B1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988. C1 Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain. D2 Two Libyan suspects were indicted in 1991.	SCU1 (w=4): two Libyans were officially accused of the Lockerbie bombing A1 [two Libyans]1 [indicted]1 B1 [Two Libyans were indicted]1 C1 [Two Libyans,]1 [accused]1 D2 [Two Libyan suspects were indicted]1	SCU2 (w=3): the indictment of the two Lockerbie suspects was in 1991 A1 [in 1991]2 B1 [in 1991]2 D2 [in 1991.]2

Figure 1: A pair of example Summary Content Units. Four semantically similar sentences from four different model summaries are grouped into two SCUs highlighting the key facts from those sentences. From Nenkova and Passonneau [61].

designed for RAG applications, metrics designed to evaluate factuality (e.g., *FactScore* [58]) or faithful manipulation of long inputs (e.g., *BooookScore* [10]) can complement application-specific evaluation frameworks.

Most approaches to automated evaluation aim to estimate the effectiveness of RAG systems across desirable dimensions (e.g., faithfulness, answer relevance, and context relevance). Techniques include prompting LLMs to evaluate generated summaries [76], and fine-tuning lightweight models on synthetic data [73]. Downstream applications, such as question answering, can also be used to evaluate the effectiveness of RAG systems [74].

4 PROPOSED FRAMEWORK

This section describes our conceptual evaluation framework for automated report generation. We name this abstract framework ARGUE (Automated Report Generation Under Evaluation) for convenience. We model the information need as a *report request*, which is analogous to the *topics* in TREC-style ad hoc retrieval evaluation. The *report writer* is required to respond with a verifiable *report*, with *citations* to its information sources. As in retrieval system evaluation, we restrict the system to citing documents in a *pre-defined document collection* instead of arbitrary information on the web. The framework is thus limited in the range of writing types it can evaluate. In particular, it does not currently support evaluation of reported information that is not explicitly supported by statements in the document collection. This restriction allows experiments that compare systems across research studies and over time.

4.1 Framework Overview

In ARGUE, creating a report generation benchmark has three phases. The first phase creates evaluation data. We believe that systems should be evaluated over human-curated data so that they are ranked on effectiveness rather than alignment to machine output.

System input comprises a document collection and report requests that describe information needs. The second phase distributes these inputs to participants. Generated reports are expected to be responsive to the information needs statements. A valid report will cite source documents that contain the reported information. Citations are a key attribute of this framework. Other report stylistic requirements might include, for example, a length limit to encourage systems to express information succinctly. If the document collection is in a language different from the report request, or is multilingual, the report may be required to be written in the language of the report request. We envision that the input data will be distributed as part of an evaluation campaign, but this is not required. Assuming an evaluation campaign, generated reports will be received and evaluated by assessors; however, to support reusability, key components will be replaced by automated mechanisms to allow future systems to be scored using the same evaluation data.

The third phase scores reports. Since the goal of this framework is to evaluate systems, each system will need to generate multiple reports based on the various report requests. Report scores will be aggregated to assign system scores. Required information in reports will be expressed by assessors in the form of nugget questions and answers. Answers will be attested in the collection and tied to particular documents that attest those answers, thereby tying the nuggets to supporting documents. During scoring, report citations will be used to determine which nuggets are described in the report. Thus there will be a notion of *recall* over nuggets, which is a new feature in RAG evaluation. Citations will also be used to ensure that non-required information that is included in the report (facts that are not part of the necessary nuggets) is attested in the collection. A *precision* score over *report segments* measures how well the report adheres to information found in the collection. This allows hallucination to be addressed, whether it be false information or true information that is unattested. While traditional recall and precision are set measures, they can be modified to account for some nuggets having greater weight than others or to allow report segments to bear multiple citations.

4.2 Evaluation Inputs and Outputs

4.2.1 *Evaluation Inputs.*

The first system input is the collection of items that will be used as source material for the retrieval task. While these items could be documents written in one or more languages, it is also possible for the items to be images, videos, audio files, or some combination. For the reminder of this paper, we will refer to the items as *documents.* Because of the importance of having citeable units, the document collection will be divided into *target elements*, which are typically documents, but can be smaller units of text such as passages, paragraphs, or sentences, depending on the needs of the evaluation. In this paper we will assume that an entire document has been selected as the target element. Segmentation into target elements should be done once and distributed with the collection to ensure that all systems are evaluated on an even footing. The document collection should include documents that contain sufficient information relevant to the desired report. Following most information retrieval-based evaluations, documents are assumed to be truthful; verifying the truthfulness of document contents is orthogonal to and beyond the scope of the framework. Instead, the framework focuses on citation, requiring that all reported information cites supporting documents from the evaluation document collection. Information that cites a document incorrectly or that is missing a required citation is appropriately penalized.

The second system input is a set of assessor-developed information needs referred to as *report requests.* A report will be generated for each report request. Report requests are more extensive and subtler than information needs for previous IR or summarization tasks. See Section 2 for the full report request description.

Creation of report requests is a complex process that tries to satisfy multiple, sometimes conflicting goals. It bears many similarities to topic creation for a TREC-style IR evaluation [85]. In topic identification for ARGUE, the topic creator must be familiar both with information retrieval, and with any special requirements of the document collection. For example, a bilingual document collection would require that the topic creator be at least bilingual. A document collection on medical topics would require topic creators who were well-versed in the medical domain.

In addition, an IR evaluation typically tries to control the number of documents that are relevant to the topic being developed, in part because doing so can improve reusability. An ARGUE evaluation must control not only the number of documents that contain relevant information, but also the number of nuggets and the number of target elements that align to each nugget. Having too many items in any of these categories leads to high assessment costs; having too few leads to higher score variance and lower ability to distinguish systems. That said, assessors need not capture all information that might satisfy the information need. It is up to the assessor to determine what, in their opinion, is the essential information.

4.2.2 Evaluation Output. The report will be generated by an automated report writer. Reports produced by the report writer should satisfy the constraints listed in Section 2. For the purposes of this framework, we make a convenience assumption that the report requester and the report audience are the same. As an example, the assessor could have the role of analyst, with the purpose of the report being to support the process of drawing analytic conclusions.

The generated report will be segmented into *report segments*, either manually or automatically. For convenience, we will assume in this work that a report segment is a sentence, but it could be some other well-defined portion of report text. Finer-grained segments may enable more nuanced distinctions. Given that precision scores operate over report segments, and given that automated sentence segmentation is imperfect, we believe that it is important that the report writer control the segmentation. Thus, each report must be segmented into sentences by the report writer prior to evaluation. The evaluation should include guidelines on sentence segmentation. The report must also include appropriate citations, pointers from source elements (sentences) to target elements (documents). Each report sentence will bear zero or more citations, as described below.

4.3 Citations

Each substantive sentence of a submitted report must cite the document target element(s) from which it was derived. Which sentences are substantive may vary according to the goals of the evaluation. A citation then is a pointer from one report segment to one target element. A given report segment may bear more than one citation, and a given target element may be cited more than once. By traversing such citations the evaluation system can map sentences in the report to documents and then to nuggets. Note that the report writer must know nothing about the nuggets that will be used to evaluate the report; they are known exclusively to the assessor. The assessor may choose to require just one citation per sentence, or, if completeness is to be measured, all valid and salient citations.

The validity of a citation has three components. First, the report segment must be supported by the target element. That is, reading the target element should verify the sentence's accuracy. In a manual evaluation, the assessor decides whether a given sentence is supported by the target element. In an automated evaluation, support of a report segment for a target element could be measured in several ways. The simplest is a semantic match, testing whether the semantics of the two texts match. A number of such automated metrics are available, such as Sentence-BERT [71]. A more accurate but harder measurement would be whether the target element entails the report sentence. Entailment has been a component of evaluation sets such as GLUE [87] and SUPERGLUE [86], and good solutions to the problem have been identified [67].

Second, at the same time, the sentence bearing the citation should be responsive to the report request. This means that the cited target element is linked to a nugget, and that the report segment provides an answer to one of that nugget's questions (see below for nugget questions). Thus the acceptability of a nugget answer depends on which document the report cites. Again, the assessor will determine whether the report segment answers a nugget question. One way to automate assessment of responsiveness might be to use an automated QA system to find answers to a nugget question, then use a semantic matching system to determine whether the report segment matches one of those answers.

Third, some evaluations will also assess whether a talented author in the field of the report would include that citation if they had written the report. An evaluation that simply wants all substantive sentences to bear a citation will omit this component; a more nuanced evaluation of reports in their final form could include it. In either case, judgments will need to be made on which sentences require a citation. Cases where no citation is required include introductory sentences, background sentences that reflect the problem statement, and sentences that summarize other cited sentences. If we are interested only in nugget recall, we can safely ignore whether sentences ought to have citations. But if we are interested in precision, we would not like to penalize a report for containing such non-citing sentences (except perhaps when measuring the quality of the report as a whole). To handle non-citing sentences, it must be determined whether the sentence should have a citation. If a citation is not needed, the report can be scored as if the identified sentences were not present in the report.

4.4 Nuggets

The proposed evaluation is centered on *nuggets.* A nugget is a piece of information that should appear in the report and that could be expressed in a variety of ways in the document collection.

4.4.1 Nugget Definition. A *nugget* in this framework is a combination of a question and one or more answers to that question that address some aspect of the report request and that are expressed in at least one target element in the collection. Nuggets must be expressed at an appropriate level of granularity for the desired report. If the report answers such a question using appropriate citations into the document collection, we deem it to have succeeded in identifying that nugget; evaluation metrics (described in Section 4.5 below) can then use statistics over the correctly answered, incorrectly answered, and unanswered nugget questions to produce a score for a given report. Answers to nugget questions should express the information that a reasonable person would expect in a report written in response to the report request.

The concept of nuggets arose from summarization evaluation [61]. New in this framework is the expression of nuggets as questions with allowable answers. We are interested in evaluation data that can be used to automatically evaluate systems, much like relevance assessments can be used to evaluate an IR system even decades after their creation. We believe this formulation will be helpful in automating report generation evaluation.

Nuggets need not capture everything any report responding to the report request might legitimately include. Given that reports by necessity will be shorter than the source documents, the assessor will determine the required information and express that as nuggets, reinforcing the idea that nuggets are opinions instead of facts. The set of answers to a nugget question are drawn from all the answers supported by the document collection. Questions and answers will be in the request language even if, for example, the source information comes from an image or is in a different language.

4.4.2 Nugget Identification. Nuggets are identified by the assessor. Nuggets must be both relevant to the report request and attested in the document collection. In practice, the assessor could either look through retrieved documents to identify important aspects of the topic from the target elements, or identify nuggets a report on the topic ought to include, then search the document collection to see which are attested. A combination of both methods could be used. To ensure reproducibility and enable evaluating recall, it is desirable to identify most (or all) nuggets that should be included.

In addition to identifying the set of nuggets for a report request, the assessor must also identify each target element in the document collection that supports an answer to each nugget. To do so, the assessor must have both a way to identify target elements that contain nugget-supporting information, and a way to bind target elements to nugget answers. The former problem is similar to that faced by many IR collection developers of ensuring that all or most relevant items have been discovered. Full collection annotation is not practical for large collections. Three main techniques for identifying relevant documents are interactive search, pooling [36, 81, 94], and active learning [14, 42, 68]. Interactive search is simply having a person use any desired tool to identify relevant documents. In pooling, the assessor judges only documents found in an aggregate of several systems' top results. Either assessors must have access to systems that together are likely to find most of the relevant documents, or this step must wait until task participants have submitted their runs. It is usually desirable to augment the pools manually using interactive search. In active learning, a classifier identifies relevant documents. Each time the assessor judges a document, the classifier is retrained to take the new judgment into account. Any or all of these techniques might be used to restrict the number of documents that must be examined during nugget identification.

The second task, assigning target elements to nuggets, is more challenging. We highlight three challenges here. First is within-nugget variation. For example, one nugget answer might be a superset of another, such as "June" versus "26 June." If the more general answer is acceptable, the more specific answer must be included in the answer set to distinguish it from an incorrect answer such as "12 June." The summarization community introduced *hoppers* [79] to capture commonality across descriptions that differ in some details. For example, two descriptions of a particular natural disaster might indicate different numbers of casualties; perhaps the descriptions were written at different times or based on two different information sources. Whether hopper-style conflation is used for a given evaluation depends on the desired report type. An overall report on the natural disaster might use hoppers; a report on how different news services covered the disaster might need to distinguish differing descriptions. As with decisions on nugget creation, if hoppers are used, the choice of hoppers is left to the assessor.

A second challenge is a single report segment or target element expressing information about more than one nugget. This is handled through multiple citations borne by a single report sentence, and/or multiple mappings between target elements and nuggets. This complicates the bookkeeping needed to give appropriate credit to each nugget, but poses no theoretical problems.

A third challenge is a single nugget requiring multiple report sentences or target elements to be fully captured. This challenge arises because nugget question/answer pairs lend themselves well to simple facts expressed in the report, but are less well suited to identifying complex information. Nonetheless we believe that the general framework will be extensible to complex nuggets whose expression is distributed across several report sentences or target elements by allowing complex questions answered by Boolean combinations of target elements, and by exploiting recent research in question answering [17, 18].

4.4.3 Practical considerations. The following considerations are not requirements of the framework, but instead practical tips we have gleaned working to instantiate this and similar evaluation frameworks. First, we believe that an assessor must be familiar both with IR concepts and any special requirements of collection and evaluation topic area (such as the aforementioned bilingual or medical settings). Second, it may be advantageous for an assessor to produce a gold standard report to help assemble the information that should be in a satisfactory report. Nugget questions can then be composed from that report. Creating a gold standard report also enables a ROUGE evaluation for comparison. Third, IR evaluations usually limit the number of relevant documents to simplify and reduce the cost of evaluation. Report evaluation would also like to control the number of nuggets and document mappings to ensure the evaluation can distinguish good and bad systems; however, this can eliminate from consideration practical use cases that would otherwise be in scope for the task. This tradeoff has traditionally been considered worthwhile, but it should be remembered that it is a tradeoff. Fourth, LLMs can call on memorized knowledge not found in the document collection. Often the LLM training collection is unknown. If the LLM has not seen the evaluation corpus, it will need to rely on hallucination, which will negatively affect evaluation data quality. Finally, while finding all potential nuggets is unnecessary

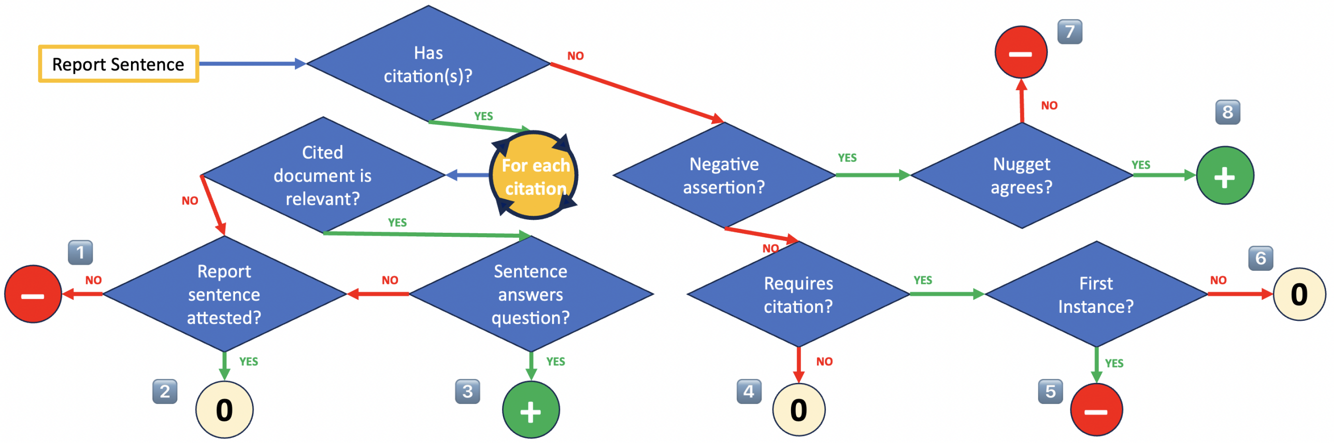

Figure 2: Report sentence scoring. Answers to eight yes/no questions dictate an outcome for each input sentence. + indicates that the sentence is rewarded, - that it is penalized, and 0 that it does not affect the overall report score.

Report Request: I am a Hollywood reporter writing an article about the highest grossing films Avengers: Endgame and Avatar. My article needs to include when each of these films was considered the highest grossing films and any manipulations undertaken to bring moviegoers back to the box office with the specific goal of increasing the money made on the film.

Gold Standard Report: Avatar originally became the highest grossing film in 2010 [D1]. Avengers: Endgame replaced Avatar as the highest grossing film in 2019 [D1, D2, D3, D8, D10, D12, D13]. It overtook Avatar by adding an additional six minutes of footage to the film to draw viewers back to the movie theater [D4]. Two years later Avatar was re-released in mainland China [D1, D2, D5, D6, D7, D8, D9, D10, D11]. It earned a sufficient amount of money to retake the title of highest-grossing film in 2021 [D5, D11, D6, D7, D2, D8, D9, D1].

Nuggets as Questions and Answers:

(1) When did Avatar first become the highest grossing film?
- 2010 [D1]

(2) When did Avengers: Endgame become the highest grossing film?
- 2019 [D1,D2, D3, D8, D10, D12, D13]
- July 2019 [D3, D12, D13]
- July 20, 2019 [D3]
- July 21, 2019 [D13][†]

(3) What did studio executives do to the Avengers: Endgame film to become the highest grossing film?
- Added six minutes of additional footage [D4]
- Added footage [D4]
- Added 6 minutes [D4]
- Additional footage at the end of the film [D14]

(4) When did Avatar retake the title of highest grossing film?
- 2021 [D1, D2, D6,D7,D9,D11]
- March 2021 [D1, D6 ,D7, D9, D11]
- March 13, 2021 [D1, D6, D9]
- Two years after the Avengers: Endgame became the highest grossing film [D2]

(5) What event led to Avatar becoming the highest grossing film?
- Re-release in Mainland China [D1, D2, D5, D6, D7, D8, D9, D10]
- Re-release in China [D1, D2, D5, D6, D7, D8, D9, D10]
- Release in Mainland China for a second time [D1, D2, D5, D6, D7, D8, D9, D10]
- Returned to theaters in China [D11]

[†] In Taiwan Time

Figure 3: Example evaluation material for a report request.

since nugget worthiness is an assessor's opinion, finding all answers to nugget questions is important for collection reusability, especially as nuggets can only be answered using documents known to have the answer. If nuggets are generated prior to submissions, it might be worth pooling submissions to identify more nugget answers.

4.5 Metrics

Many metrics can be used to assess automatically generated reports. Two common IR measures are recall and precision; we focus on these here because they are well-known, easy to calculate, and highlight most of the important scoring issues we face in generated report evaluation. Recall and precision each require a numerator and a denominator. The recall denominator is the number of distinct assessor-identified nuggets; its numerator is the number of correctly reported nuggets (those supported by one or more of the necessary supporting citations in the report). So recall tells us how many of the concepts central to the report were actually reported on. Precision must account for phenomena below the nugget level, so we calculate it over report segments (which again we assume to be sentences). The denominator is the number of report sentences, minus any sentence that does not require a citation or that properly cites information not part of any nugget. The numerator is the number of sentences deemed to bear accurate citations, plus any sentences specified by the evaluation that correctly bear no citation.

Figure 2 describes a typical approach to sentence evaluation. The rules embedded in the flowchart are not hard-and-fast, but are likely adequate for many evaluations under this framework. In the flowchart, "+" means the sentence is correct and should be rewarded; "–" means that it is incorrect and should be penalized; and "0" means that the sentence is not included in the score. The flowchart shows how each sentence of the report can be scored. We propose these principles to guide sentence scoring:

- Sentences with citations whose target element does not support them should be penalized (Outcome#1 in Figure 2).
- Properly cited and attested sentences that are not relevant to the report should be ignored (Outcome #2).

(1) Avengers: Endgame and Avatar are two of the highest-grossing films in history. Outcome: #4
(2) Avengers: Endgame surpassed Avatar as the highest-grossing film globally, with a box office revenue of $2.787 billion. [D12] Outcome: #2
(3) This achievement was officially announced in July 2019. [D13] Outcome: #3, Nugget: 2
(4) To bring moviegoers back to the box office and increase the revenue of Avengers: Endgame, Marvel Studios implemented several strategies. Outcome: #4
(5) One of these strategies was re-releasing the film with additional content. Outcome: #6
(6) Marvel Studios President Kevin Feige confirmed plans for a re-release, including previously unseen footage, to generate renewed interest and boost ticket sales. [D101] Outcome: #2
(7) The new version of the film did not include any re-edited scenes within the main feature but instead featured additional scenes after the end credits, including a tribute. [D14] Outcome: #3, Nugget: 3
(8) This approach aimed to entice both existing and new audiences to revisit the theaters and contribute to the film's box office success. Outcome: #6
(9) Similarly, Avatar also made a comeback to reclaim its position as the highest-grossing film. Outcome: #4
(10) The film was re-released in China, where it achieved a single-day box office revenue of $8 million. [D2] Outcome: #3, Nugget: 5
(11) The re-release in China was particularly successful, as it generated significant revenue and propelled Avatar back to the top of the global box office charts. [D1] Outcome: #3, Nugget: 5
(12) The film's re-release in China was attributed to the impact of the COVID-19 pandemic, which led to the delay of many new releases and created an opportunity for Avatar to regain its popularity. [D2] Outcome: #3, Nugget: 5
(13) In conclusion, Avengers: Endgame and Avatar both held the title of the highest-grossing film globally at different points in time. Outcome: #4
(14) Marvel Studios strategically re-released Avengers: Endgame with additional content to attract audiences and boost ticket sales. Outcome: #6
(15) Avatar capitalized on the re-release trend in China, taking advantage of the pandemic-induced delay of new releases. Outcome: #6
(16) These manipulations aimed to increase the films' box office revenue and solidify their positions as record-breaking blockbusters. Outcome: #4

Figure 4: Example report evaluation result.

- A sentence that cites a target element supporting a nugget that the sentence fulfills should be rewarded (Outcome #3).
- Sentences that neither have nor require citations should not affect the score (Outcome #4).
- Sentences that should contain a citation but do not should be penalized the first time their claim occurs (Outcomes #5, #6).
- Sentences that claim the absence of a fact should be rewarded or penalized depending on whether the absence is explicitly stated as a nugget (Outcomes #7, #8). For this, a nugget can be created for information that the report request explicitly asks for but is not attested in the collection.

Most sentences will bear either zero or one citation. A sentence can bear multiple citations, either because the same information is multiply attested in the collection, or because it is complex. Sentences that cite multiple target elements supporting the same nugget are treated as a single citation. Alternatively, the evaluation may macroaverage citation scores if all sentences are to be given equal weight, or microaverage them if the focus is on citation accuracy. Support by multiple report sentences counts only once per nugget.

To automatically score a report, each decision diamond in Figure 2 must be automatable. Some are trivial, such as "Has citation;" others are less so. We believe current technology could do a reasonable job with most of the tasks. For instance, entailment models can likely determine if a document supports a report sentence. Note that originality is not a component of this evaluation; preventing plagiarism, while important, is a specialized area with its own metrics and evaluations [5, 22, 34, 66].

5 EXAMPLE ASSESSMENT

Figure 3 shows an example of the two items required to do manual or automatic assessment. The report request identifies the desired report content. The nugget questions and answers show how each answer is linked to the documents that attest to that answer. The Gold Standard Report that is shown is optional, but a useful intermediate step for the assessor between source document search and nugget question creation.

Figure 4 is a report generated in response to the example in Figure 3, broken into report segments to illustrate manual evaluation. Each Outcome: # indicates how the sentence would be categorized using the flowchart in Figure 2. For Outcome: #3, the nugget answer in the sentence is also recorded. No sentence received a negative assessment because there were no outcomes of #1 or #7. Therefore, precision is $5/(16-11) = 1.0$. One nugget was repeated in Lines 10, 11, and 12, so recall is $3/5 = 0.6$. For both Lines 2 and 6, the assessor would have needed to refer to the original source document to assess the statement, since the information in the sentence had not been captured in a required nugget. Assessing such sentences will likely be the most time-consuming part of manual assessment.

6 CONCLUSIONS

LLMs have enabled remarkable new ways to satisfy information needs. Rather than simply providing "10 blue links" or an extracted answer snippet, LLMs have the potential to peer into documents to identify information salient to a topic and compile it into highly coherent, long-form text responses. We envision these generated reports will be a central way that some users will satisfy complex, nuanced, or multifaceted information needs.

Because we believe that current evaluation methodologies for these report-generation systems are insufficient to maintain quality and guard against known defects, we felt the need for a report evaluation framework based on core principles — responsiveness to the information need, grounding and verifiability in documents, completeness, and reusability — while deliberately omitting aspects of report generation that current systems do not seem to struggle with (e.g., coherence, structure, etc.). Our new perspective on report generation evaluation is IR-centric, pulling together tried-and-true notions of relevance, recall, and user modeling. We have also demonstrated an instantiation of our framework that could be applied either manually or with automatic systems.

Evaluation methodologies inform progress and direct attention. We hope our proposed generated report evaluation framework will spur progress in the development of next-generation information access systems that can provide responsive, complete, and verifiable information on complex, nuanced, and multifaceted topics.

WHOSE PERSPECTIVE

This paper represents the perspectives of a group of industry and academic researchers at a variety of career stages. We thank John Conroy, Hoa Dang, Mohit Iyyer, Jimmy Lin, and Graham Neubig for their discussions with us.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. 2021. RetrievalSum: A retrieval enhanced framework for abstractive summarization. *arXiv preprint arXiv:2109.07943* (2021).

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390 [cs.CV]

[4] Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, et al. 2023. MegaWika: Millions of reports and their sources across 50 diverse languages. *arXiv preprint arXiv:2307.07049* (2023).

[5] Anton Belyy, Marina Dubova, and Dmitry Nekrasov. 2018. Improved Evaluation Framework for Complex Plagiarism Detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 157–162. https://aclanthology.org/P18-2026

[6] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. *ArXiv* abs/2212.08037 (2022). https://api.semanticscholar.org/CorpusID:254685584

[7] Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *57th Annual Meeting of the Association-for-Computational-Linguistics (ACL)*. 1800–1809.

[8] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. *arXiv preprint arXiv:2105.11269* (2021).

[9] Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 152–161.

[10] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *ArXiv* abs/2310.00785 (2023). https://api.semanticscholar.org/CorpusID:263605928

[11] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. *ArXiv* abs/2309.01431 (2023). https://api.semanticscholar.org/CorpusID:261530434

[12] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928* (2022).

[13] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*. 93–98.

[14] Gordon V. Cormack, Haotian Zhang, Nimesh Ghelani, Mustafa Abualsaud, Mark D. Smucker, Maura R. Grossman, Shahin Rahbariasl, and Amira Ghenai. 2019. Dynamic Sampling Meets Pooling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1217–1220. https://doi.org/10.1145/3331184.3331354

[15] Hoa Trang Dang, Diane Kelly, Jimmy Lin, et al. 2007. Overview of the TREC 2007 Question Answering Track.. In *Trec*, Vol. 7. 63.

[16] Márcio de S. Dias, Valéria D. Feltrim, and Thiago Alexandre Salgueiro Pardo. 2014. Using Rhetorical Structure Theory and Entity Grids to Automatically Evaluate Local Coherence in Texts. In *Computational Processing of the Portuguese Language*, Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes (Eds.). Springer International Publishing, Cham, 232–243.

[17] Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive Prompting for Decomposing Complex Questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1251–1265. https://doi.org/10.18653/v1/2022.emnlp-main.81

[18] Romina Etezadi and Mehrnoush Shamsfard. 2022. The state of the art in open domain complex question answering: a survey. *Applied Intelligence* 53 (06 2022). https://doi.org/10.1007/s10489-022-03732-9

[19] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. *arXiv preprint arXiv:1907.09190* (2019).

[20] Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. *arXiv preprint arXiv:2203.10299* (2022).

[21] Rafael Ferreira Mello, Giuseppe Fiorentino, Hilário Oliveira, Péricles Miranda, Mladen Rakovic, and Dragan Gasevic. 2022. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 404–414.

[22] Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Comput. Surv.* 52, 6, Article 112 (oct 2019), 42 pages.

[23] Tianyu Gao, Ho-Ching Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. *ArXiv* abs/2305.14627 (2023). https://api.semanticscholar.org/CorpusID:258865710

[24] Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated Pyramid Summarization Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Mohit Bansal and Aline Villavicencio (Eds.). Association for Computational Linguistics, Hong Kong, China, 404–418. https://aclanthology.org/K19-1038

[25] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Frobe, Guide Zucoon, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. Evaluating Generative Ad Hoc Information Retrieval. *ArXiv* abs/2311.04694 (2023). https://api.semanticscholar.org/CorpusID:265050661

[26] John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2023. Open Domain Multi-document Summarization: A Comprehensive Study of Model Brittleness under Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 8177–8199.

[27] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *ArXiv* abs/2209.12356 (2022). https://api.semanticscholar.org/CorpusID:252532176

[28] Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 128–137. https://aclanthology.org/D15-1013

[29] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview.. In *TREC*.

[30] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *ArXiv* abs/2002.08909 (2020). https://api.semanticscholar.org/CorpusID:211204736

[31] Donna Harman and Paul Over. 2004. The Effects of Human Variation in DUC Summarization Evaluation. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 10–17. https://aclanthology.org/W04-1003

[32] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3905–3920. https://aclanthology.org/2022.naacl-main.287

[33] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23369–23379.

[34] Kamal Mansour Jambi, Imtiaz Hussain Khan, and Muazzam Ahmed Siddiqui. 2022. Evaluation of Different Plagiarism Detection Methods: A Fuzzy MCDM Perspective. *Applied Sciences* 12, 9 (2022). https://www.mdpi.com/2076-3417/12/9/4580

[35] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446.

[36] Sabrina Keenan, Alan F. Smeaton, and Gary Keogh. 2001. The effect of pool depth on system evaluation in TREC. *Journal of the Association for Information Science and Technology* 52, 7 (May 2001), 570–574. https://doi.org/10.1002/asi.1096.abs

[37] Judith L Klavans, Min-yen Kan, and Kathleen McKeown. 2001. Domain-specific informative and indicative summarization for information retrieval. In *SIGIR Worshop on Text Summarization*.

[38] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566* (2021).

[39] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. *arXiv preprint arXiv:2103.06332* (2021).

[40] Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (Edinburgh, Scotland) *(IJCAI'05)*. Morgan

Kaufmann Publishers Inc., San Francisco, CA, USA, 1085–1090.

[41] Patrick 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. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[42] Dan Li and Evangelos Kanoulas. 2017. Active Sampling for Large-scale Information Retrieval Evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3132847.3133015

[43] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110* (2022).

[44] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437* (2018).

[45] Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6191–6196.

[46] 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://aclanthology.org/W04-1013

[47] Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 457–464.

[48] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 150–157. https://aclanthology.org/N03-1020

[49] Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809* (2022).

[50] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 997–1006. https://aclanthology.org/P11-1100

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV]

[52] Jerry Liu. 2022. *LlamaIndex*. https://github.com/jerryjliu/llama_index

[53] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. *ArXiv* abs/2304.09848 (2023). https://api.semanticscholar.org/CorpusID:258212854

[54] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:254877603

[55] Inderjeet Mani. 2001. Automatic summarization. *Automatic Summarization* (2001).

[56] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2014. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 301–310.

[57] Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2023. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. *ArXiv* abs/2308.04430 (2023). https://api.semanticscholar.org/CorpusID:260704206

[58] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *ArXiv* abs/2305.14251 (2023). https://api.semanticscholar.org/CorpusID:258841470

[59] Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2016. Overview of TAC KBP 2016 Event Nugget Track.. In *TAC*.

[60] Gianluca Moro, Stefano Salvatori, and Giacomo Frisoni. 2023. Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval. *Neurocomputing* 538 (2023), 126196. https://www.sciencedirect.com/science/article/pii/S092523122300303X

[61] Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 145–152. https://aclanthology.org/N04-1019

[62] Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2021. QuALITY: Question Answering with Long Input Texts, Yes! *arXiv preprint arXiv:2112.08608* (2021).

[63] Jong Won Park. 2020. Continual BERT: Continual learning for adaptive extractive summarization of COVID-19 literature. *arXiv preprint arXiv:2007.03405* (2020).

[64] Hao Peng, Ankur P Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. *arXiv preprint arXiv:1904.04428* (2019).

[65] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *NeurIPS*.

[66] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Coling 2010: Posters*, Chu-Ren Huang and Dan Jurafsky (Eds.). Coling 2010 Organizing Committee, Beijing, China, 997–1005. https://aclanthology.org/C10-2115

[67] I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2023. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express* (2023). https://www.sciencedirect.com/science/article/pii/S2405959523001145

[68] Md Mustafizur Rahman, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2020. Efficient Test Collection Construction via Active Learning. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 177–184. https://doi.org/10.1145/3409256.3409837

[69] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250* (2016).

[70] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics* (2023), 1–64.

[71] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084* (2019).

[72] Adam Roegiest, Gordon V Cormack, Charles LA Clarke, and Maura R Grossman. 2015. TREC 2015 Total Recall Track Overview.. In *TREC*.

[73] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476* (2023).

[74] 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 *Biennial Conference on Design of Experimental Search & Information Retrieval Systems*. https://api.semanticscholar.org/CorpusID:238207962

[75] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[76] ES Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *ArXiv* abs/2309.15217 (2023). https://api.semanticscholar.org/CorpusID:263152733

[77] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567 [cs.CL]

[78] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (Lisbon, Portugal) *(CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 623–632.

[79] Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, and Christopher Caruso. 2018. Cross-document, cross-language event coreference annotation using event hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[80] Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 228–235.

[81] Karen Sparck Jones and C.J. Van Rijsbergen. 1975. *Report on the Need for and Provision of an 'ideal' Information Retrieval Test Collection*. University Computer Laboratory.

[82] Nandan Thakur, Luiz Bonifacio, Xinyu Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy J. Lin. 2023. NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation. *ArXiv* abs/2312.11361 (2023). https://api.semanticscholar.org/CorpusID:266359301

[83] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management* 43, 6 (2007), 1606–1618.

[84] Ellen M Voorhees. 2001. The TREC question answering track. *Natural Language Engineering* 7, 4 (2001), 361–378.

[85] Ellen M. Voorhees and Donna Harman. 1998. The Text REtrieval Conferences (TRECs). In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*. Association for Computational Linguistics, Baltimore, Maryland, USA, 241–273. https://aclanthology.org/X98-1031

[86] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv preprint 1905.00537* (2019).

[87] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In the Proceedings of ICLR.

[88] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. *arXiv preprint arXiv:2305.18201* (2023).

[89] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561* (2022).

[90] Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:258587884

[91] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675* (2019).

[92] Yijie Zhou, Kejian Shi, Wencai Zhang, Yixin Liu, Yilun Zhao, and Arman Cohan. 2023. ODSum: New Benchmarks for Open Domain Multi-Document Summarization. *arXiv preprint arXiv:2309.08960* (2023).

[93] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize Before You Write: Imagination-Guided Open-Ended Text Generation. *arXiv preprint arXiv:2210.03765* (2022).

[94] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 307–314. https://doi.org/10.1145/290941.291014