Measuring Attribution in Natural Language Generation Models

Hannah Rashkin*♣♢
Google DeepMind
hrashkin@google.com

Vitaly Nikolaev*♣♠
Google DeepMind
vitalyn@google.com

Matthew Lamm♠
Google DeepMind
mrlamm@google.com

Lora Aroyo♠
Google Research
loraa@google.com

Michael Collins♠
Google DeepMind
mjcollins@google.com

Dipanjan Das♠♡
Google DeepMind
dipanjand@google.com

Slav Petrov♡
Google DeepMind
slav@google.com

Gaurav Singh Tomar♢
Google DeepMind
gtomar@google.com

* Equal contribution. All authors contributed to all parts of the article. ♠ Led development of the conceptual framework. ♣ Led human annotation study. ♢ Contributed to modeling experiments. ♡ Provided project leadership and management.

Action Editor: Myle Ott. Submission received: 14 July 2022; revised version received: 17 March 2023; accepted for publication: 10 May 2023.

https://doi.org/10.1162/coli_a_00486

Iulia Turc♢
Storia AI
iulia@iuliaturc.com

David Reitter♠♡
Google DeepMind
reitter@google.com

Large neural models have brought a new challenge to natural language generation (NLG): It has become imperative to ensure the safety and reliability of the output of models that generate freely. To this end, we present an evaluation framework, Attributable to Identified Sources *(AIS), stipulating that NLG output pertaining to the external world is to be verified against an independent, provided source. We define AIS and a two-stage annotation pipeline for allowing annotators to evaluate model output according to annotation guidelines. We successfully validate this approach on generation datasets spanning three tasks (two conversational QA datasets, a summarization dataset, and a table-to-text dataset). We provide full annotation guidelines in the appendices and publicly release the annotated data at* `https://github.com/google-research-datasets/AIS`.

1. Introduction

Large, pretrained neural models have advanced Natural Language Generation (NLG) performance across a variety of use cases, including text summarization, translation, and dialogue. Yet, generative neural models are known to hallucinate often, lacking faithfulness to underlying sources—for example, in summarization or in grounded dialogue systems. Accurate evaluation with respect to these issues is important.

In this article, we develop a framework for the evaluation of **attribution**, by which we mean the accurate use of source documents to support generated text. Attribution is closely related to issues of hallucination and faithfulness (see §2 for discussion). As a key motivating example, consider a dialogue with a system that generates responses to a user's sequence of questions:

USER: what was George Harrison's first solo album?
SYSTEM: it was "Wonderwall Music", released in November 1968.
USER: how old was he when it was released?
SYSTEM: he was 25 years old

If such a system, in addition to generating responses, could attribute its statements to source documents, that is, provide sufficient and concise evidence that acts as corroboration for its claims, system designers and users alike could more readily ascertain the extent to which the information it provides is supported by underlying sources. Prior work in NLG spanning diverse use cases such as summarization, dialogue response generation, and data-to-text generation has investigated issues of faithfulness and "hallucination", but has not provided a uniform and formally expressed framework to measure these errors. We discuss the relationship of our work to related work in §2.

In §3, we introduce our evaluation framework, Attributable to Identified Sources (AIS), that can be used to assess whether statements in natural language made by a system are corroborated by a given underlying source. The definition of AIS (see

§3.3.1) formalizes the meaning of a sentence s in context using the notion of explicatures (Carston 1988; Wilson and Sperber 2004),[1] and defines attribution to some background information source P in terms of an intuitive test, asking whether "According to P, s". It also accommodates system outputs whose meaning is uninterpretable. AIS can be used as a pre-condition or in tandem with other metrics or evaluation frameworks to assess overall quality. For example, characteristics of the underlying source (such as "source quality"), the fluency of the generated text, and so forth, can be measured using complementary metrics that are out of scope in this work.

We propose specific instantiations of AIS for three NLG tasks (§4): response generation in a conversational QA setting (as in the example above; responses must be attributable to a provided answer document), text summarization (where the summary must be attributable to the source article), and description generation from structured tables, or table-to-text (where the description must be attributable to the source table and associated metadata). Each domain involves a number of challenges: For example, in dialogue systems a key challenge is that the meaning of system responses is highly contextually dependent.

Next, we establish the feasibility of AIS evaluations via an empirical study through conducting human evaluation experiments. We train annotators to evaluate output text from multiple models per task using task-specific instantiations of AIS. We show that in our human evaluation studies, it is possible to achieve a moderate to high degree of inter-annotator agreement (see §4 for more details). We're also able to observe differences in model outputs' AIS scores, following generally expected trends. As part of this work, we release the detailed guidelines for human evaluation along with annotated data. We believe that AIS as a framework would be essential for the evaluation of system-generated utterances across NLG tasks.

2. Background

Hallucinations in NLG. As alluded to in §1, past work has identified the issue of hallucination in neural generation models. Wiseman, Shieber, and Rush (2017) presented challenges in data-to-text generation where neural models generate hallucinated content not supported by source data; they proposed an automatic information extraction-based metric to evaluate generated text for that particular scenario and conducted a small human evaluation study examining whether summaries are supported by source data. More recently, Parikh et al. (2020) conducted a larger human evaluation study in the context of a data-to-text generation dataset entitled ToTTo, and measured hallucinations in terms of *faithfulness* with respect to a source data table. In text summarization, Maynez et al. (2020) characterized hallucinations in detail and discussed behavior of models that generate content that are present in larger corpora beyond a given source and conducted a significant human study. Additional automatic QA-based methods for detecting hallucinations have been proposed by Wang, Cho, and Lewis (2020), and Nan et al. (2021), among others. Perhaps most relevantly, Durmus, He, and Diab (2020) involved both a human evaluation and the introduction of an automatic question–answer based evaluation method. Their human evaluation of summary sentences is similar to our two-stage annotation pipeline where they evaluate sentences in two

1 For example, in the above dialogue the explicature of "he was 25 years old"is "George Harrison was 25 years old when "Wonderwall Music" was released": the latter explicature is evaluated for attribution. Note that this use of explicatures is closely related to prior work on decontextualization (Choi et al. 2021). See §2 for more discussion.

Example E1
u_1: what was George Harrison's first solo album?
s_1: it was "Wonderwall Music", released in July 2006.
u_2: how old was he when it was released?
s_2: 25 years old
$E(c_{2,1}, s_{2,1}) =$ *George Harrison was 25 years old when "Wonderwall Music" was released*

Example E2
u_1, s_1, u_2 as in Example E1
s_2: he was 25 years old
$E(c_{2,1}, s_{2,1}) =$ *George Harrison was 25 years old* when "Wonderwall Music" was released

Example E3
u_1, s_1, u_2 as in Example E1
s_2: He was 25 years old. It was the first solo album by a member of the Beatles.
$E(c_{2,2}, s_{2,2}) =$ *"Wonderwall Music" was the first solo album by a member of the Beatles.*

Example E4
u_1, s_1, u_2 as in Example E1
s_2: the band was The Beatles
$E(c_{2,1}, s_{2,1}) =$ `NULL`

Example E5
u_1, s_1, u_2 as in Example E1
s_2: it was 25
$E(c_{2,1}, s_{2,1}) =$ `NULL`

Example E6
u_1, s_1, u_2 as in Example E1
s_2: He was 25 years old. Have you heard that album?
$E(c_{2,2}, s_{2,2}) =$ *Have you heard the album "Wonderwall Music"?*

Example E7
u_1: what was George Harrison's first solo album?
s_1: it was "Wonderwall Music", released in July 2006.
u_2: how old was he when it was released?
$E(c'_2, u_2) =$ *how old was George Harrison when "Wonderwall Music" was released?*

Figure 1
Examples of utterances in context, and their explicatures.

steps—first for whether it is understandable, and second, if so, for faithfulness to the underlying source (their instructions to annotators are: "If the information conveyed by the sentence is not expressed in the source, select 'unfaithful'.").

In the case of response generation for dialogues, research has investigated measuring the responses' consistency to prior conversational history or their groundedness to some external evidence, which we deem to be very close to the topic of hallucination. These have been measured via dialogue-specific natural language inference methods, often via human studies and data creation (e.g., Welleck et al. 2019; Mehri and Eskenazi 2020; Gupta et al. 2022; Honovich et al. 2021; Dziri et al. 2022; Santhanam et al. 2021).

General-purpose benchmarking across these tasks have gained traction (Gehrmann et al. 2021), but there has not been a standardized treatment of the attribution problem. The research reported in this article attempts to address this gap by explicitly formalizing the evaluation of attribution as a replicable and extendable conceptual framework. As part of our definition of attribution, we outline a more formal background for "information conveyed by the text"—in particular, through the use of explicatures (see Figure 1 for examples). Lastly, we demonstrate that AIS can be generalized across

multiple NLG tasks in which context, source documents, and generated text can take different forms.

Fact Verification. A related field of study has dealt with the topic of fact or claim verification (Thorne et al. 2018; Thorne and Vlachos 2018; Thorne et al. 2021, inter alia). Work in this area has framed the task as retrieving supporting evidence given a claim, and optionally classifying semantic relationships between the claim and the evidence text. Modeling approaches have overlapped with recent literature examining natural language inference (Nie, Chen, and Bansal 2019). Thorne et al. (2021) have examined several human annotation tasks for the above family of problems; however, there are several key differences with this work. First, we evaluate the quality of a system generated utterance with respect to given evidence source (a fundamentally different end goal); we utilize the notion of explicatures in defining attribution; finally, we avoid absolute judgments regarding "factuality" of utterances. As mentioned in §1, rather than making factuality judgments, we deem that complementary evaluation methods such as "source quality" in tandem with AIS would be required to evaluate the factuality of utterances. As a corollary, we assume the source is a reference, and that an actual system may select sources for their trustworthiness.

Decontextualization. Choi et al. (2021) introduce the task of decontextualization, that is, the problem of taking a sentence in context and rewriting it in a way that its meaning is preserved, and it can be interpreted out of context. This is directly related to the idea of explicatures, which are also used in this article.

3. A Formal Definition of Attributable to Identified Sources

This section gives a formal definition of AIS, attempting to give a clear and precise definition of attribution. Having a firm formal foundation for AIS is important, in developing a set of linguistically sound definitions before developing guidelines for annotators. We will describe at the end of the section how it is relatively straightforward to convey these concepts in an intuitive way that is accessible to raters with or without a background in linguistics.

We first give a definition of AIS for a simple case, where the utterance from a system is a standalone proposition. In spite of the simplicity of this setting, it is highly informative, and forms the basis for the full definition of AIS. We then describe how this definition extends to a much larger set of system utterances, in particular giving a treatment of *interpretability*[2] and *contextual effects*. A key idea in our model of meaning in context is the notion of explicatures (Carston 1988; Wilson and Sperber 2004; Choi et al. 2021). In a final subsection, we describe how key aspects of the AIS definition naturally lend it to operationalization.

2 We acknowledge that the term "interpretability" has come to signify "model interpretability" in the NLP and ML communities (as established in Harrington et al. [1985], Ribeiro, Singh, and Guestrin [2016]). The term in our use represents how interpretable system output is for a human annotator. The choice of terminology is intended to be more conceptually transparent when used by annotators: Unlike other terms like "meaningful"/"nonsensical" (Durmus, He, and Diab 2020), or "sensibleness" (Adiwardana et al. 2020), "interpretability" more readily alludes to the significance of the propositions in system generated output in relationship to context. Finally, the annotators are typically not familiar with the "model interpretability" usage of the term.

3.1 An Initial Definition of AIS: Attribution of Standalone Propositions

We now give a definition of AIS for a simple but important case, where the text in question is a *standalone proposition*. We in general assume a setting where AIS is to be determined for a string whose meaning is ascertained relative to a context. In the following treatment we assume that time is the only non-linguistic aspect of context relevant to determining textual meaning, modeling a setting where two generic speakers communicate over a text-based channel, with no additional prior information about each other. Extensions of AIS to more complex settings may require a more elaborate notion of non-linguistic context.

We define standalone propositions as follows:

Definition 1 (Standalone Propositions)
A standalone proposition is a declarative sentence that is interpretable once a time of interpretation t has been specified.

To illustrate the definition of standalone propositions, consider the following examples:

Example S1: George Harrison was 25 years old when his album "Wonderwall Music" was released.
Example S2: He was 25 years old.
Example S3: George Harrison was 25 years old.
Example S4: George Harrison died over 15 years ago.

All four examples are declarative sentences. S1 is a standalone proposition. S4 is a standalone proposition, as it is interpretable once the time t is specified. S2 is, however, not a standalone proposition, as it cannot be interpreted without additional contextual information: It is unclear what "He" refers to. More subtly, S3 is also not a standalone proposition, because it lacks a temporal point of reference.

The definition of AIS for standalone propositions is as follows:

Definition 2 (AIS for Standalone Propositions)
A pair (s, t) consisting of a standalone proposition s and a time t is Attributable to Identified Sources (AIS) iff the following conditions hold:

1. The system provides a set of parts P of some underlying corpus K, along with s.
2. (s, t) is attributable to P.

A pair (s, t) is **attributable** to a set of parts P of some underlying corpus K iff: A generic hearer will, with a chosen level of confidence, affirm the following statement: "According to P, s", where s is interpreted relative to time t.

Here, the corpus K could be a set of web pages, and the parts P could be pointers to paragraphs or sentences within K; or the corpus K could be a knowledge graph, with P as parts of the underlying knowledge graph; other examples are no doubt possible.

Remark: In defining AIS we use the idea of the generic hearer, which is a theoretical idealization, for convenience. Actual AIS judgments will be computed by sampling from a pool of raters, who may differ in the amount of background knowledge they

Table 1
Examples illustrating the complexities in AIS judgments. These types of examples require extra reasoning, or assumptions about shared background knowledge. The examples are purely illustrative (not from real data examples).

Evidence	Proposition Candidate	Challenges
George Harrison (25 February 1943—29 November 2001) was an English musician…His debut solo album was "Wonderwall Music", released in November 1968.	George Harrison was 25 years old when his album "Wonderwall Music" was released.	*Common sense and cultural knowledge is required to interpret the information in the proposition as it requires inferring that "his" is still referring to George Harrison. "George" is typically a male name in English; musicians release albums; therefore, "his album" likely refers to George Harrison, but not another unattested entity.*
The runtime of the theatrical edition of "The Fellowship of the Ring" is 178 minutes, the runtime of "The Two Towers" is 179 minutes, and the runtime of "The Return of the King" is 201 minutes.	The full run-time of "The Lord of the Rings" trilogy is 558 minutes.	*Evaluating this requires numerical reasoning, and it also requires knowing that "The Lord of the Rings" trilogy consists of the three films mentioned (background knowledge that may vary from person to person). Additionally, it requires assumptions that the runtime is exclusively referring to the theatrical edition of these movies.*

bring to the task, or in the strictness with which they apply the "according to" test (see Table 1 for examples of the background knowledge and reasoning involved in making AIS judgments).

As an example, consider standalone proposition S1 given above, assume that the corpus K is all of Wikipedia, t_0 is the present time (specifically, noon on December 21, 2021), and assume that the set P consists of a single paragraph from Wikipedia, as follows:

Example P1: George Harrison (25 February 1943–29 November 2001) was an English musician, singer–songwriter, and music and film producer who achieved international fame as the lead guitarist of the Beatles. His debut solo album was "Wonderwall Music", released in November 1968.

Under this definition, it would be correct for a hearer to judge "$(S1, t_0)$ is attributable to P1", because the "according to" test in the AIS definition holds. That is, it is reasonable to say "according to P1, S1" where S1 is interpreted at time t_0: "according to P1, George Harrison was 25 years old when his album "Wonderwall Music" was released."

Note that in some cases the system may provide multiple parts. The standalone proposition S may also be justified by certain forms of multi-hop reasoning (e.g., arithmetic processes) over that set of parts. For instance, the above example requires reasoning about dates and age.

3.2 Extending AIS: Attribution of Sentences in Context

We now extend the previous definition of AIS to cover sentences that go beyond standalone propositions. To do so, we will need to consider multi-sentence cases, and cases with non-empty linguistic contexts. We will also cover cases that are uninterpretable.

We first define the notion of "utterance":

Definition 3 (Utterance)
An **utterance** is a sequence of one or more *sentences* produced by a system or user, where a sentence may be a declarative, a question, a command, an exclamation, or a fragment. The ith system utterance is $s_i = s_{i,1} \dots s_{i,|s_i|}$, where $s_{i,j}$ is the jth sentence within system utterance s_i, and similarly the ith user utterance is $u_i = u_{i,1} \dots u_{i,|u_i|}$.

To briefly illustrate our approach to non-empty linguistic contexts, consider the following interaction between a user and system (originally given in the Introduction; repeated here for convenience):

u_1**:** what was George Harrison's first solo album?
s_1**:** it was "Wonderwall Music", released in November 1968.
u_2**:** how old was he when it was released?
s_2**:** he was 25 years old.

The system utterance s_2 = *he was* 25 *years old* is clearly not a standalone proposition. As such, it cannot be evaluated for AIS given our previous definition. However, given the previous context in the interaction, intuitively the meaning of s_2 is something similar to the standalone proposition "George Harrison was 25 years old when his album "Wonderwall Music" was released". This latter "paraphrase" of s_2's meaning is a standalone proposition, and can be evaluated using the AIS definition for standalone propositions.

We will make this notion of "paraphrase" of the meaning of an utterance in context more formal, through the introduction of *explicatures*. The explicature of s_2 in context of the previous utterances u_1, s_1, u_2 is *e = George Harrison was 25 years old when his album "Wonderwall Music" was released*. Once explicatures have been defined in this way, they can be evaluated for AIS in exactly the same way as standalone propositions.

3.2.1 Definition of Interactions and Linguistic Context. We will use the following definition of **interaction** throughout the article:

Definition 4 (Interactions)
An *interaction* consists of: (1) a sequence $u_1 \dots u_m$ of $m \geq 0$ user utterances; (2) a sequence $s_1 \dots s_n$ of $n \geq 0$ system utterances; and (3) a strict total order over the $m + n$ user and system utterances.[3]

This setting is intended to be quite general, including a broad class of applications where systems generate utterances. In conversational QA systems we typically have alternating user and system utterances, where $m = n$, and the total ordering is $u_1, s_1, u_2, s_2, \dots u_n, s_n$. In summarization tasks we have a simplified setting where $m = 0$, $n = 1$, and s_1 is equal to the summary generated by the system. Table-to-text tasks are

3 For example, the order might be specified by functions $U : \{1 \dots m\} \rightarrow \{1 \dots (m + n)\}$ and $S : \{1 \dots n\} \dots (m + n)\}$ where $U(i)$ (respectively, $S(i)$) is the position of utterance u_i (respectively, s_i) in the total ordering. The notational details will not be important for this article.

similar to summarization in that $m = 0$, $n = 1$, while s_1 is the description of the table generated by the system.

Each sentence has an associated linguistic context:

Definition 5 (Linguistic Context for Sentences)
We define the linguistic context for system sentence $s_{i,j}$ to be $c_{i,j}$, where $c_{i,j}$ is the ordered sequence of sentences (with speaker identities, user or system) that precedes $s_{i,j}$ in the total ordering. We define the linguistic context for user sentence $u_{i,j}$ to be $c'_{i,j}$, where c' is defined in a similar way.[4]

Here the definition of "sentence" is intended to be quite broad. A sentence could be a declarative sentence, a question, or a fragment (such as the string "25 years old"). Under the above definitions, the context for a user or system sentence is simply the sequence of user and system sentences that precedes it. To illustrate these definitions consider the following example:

u_1: what was George Harrison's first solo album?
s_1: it was "Wonderwall Music" released in November 1968.
u_2: how old was he when it was released?
s_2: He was 25 years old. It was the first solo album by a member of the Beatles.

Here the system utterance s_2 consists of two sentences, $s_{2,1} =$ *He was 25 years old* and $s_{2,2} =$ *It was the first solo album by a member of the Beatles*.

3.3 Explicatures

A key goal in this section is to define AIS for sentences $s_{i,j}$ in linguistic contexts $c_{i,j}$ that are non-empty (i.e., which contain previous sentences in the discourse). To do this it will be critical to formalize what is meant intuitively by "the meaning of $s_{i,j}$ in context $c_{i,j}$". To do this we introduce **explicatures** (this definition is closely related to definition 1 in Choi et al. [2021]):

Definition 6 (Explicatures)
Define the context c to be (c_l, t), where c_l is the linguistic context and t is the time. Define $\bar{c}$ to be the context (ϵ, t) where ϵ is the linguistically empty context: that is, $\bar{c}$ is a copy of c but with c_l replaced by ϵ. The set of *explicatures* $E(c, x)$ of a sentence x in a context c is a set that satisfies the following conditions: (1) each $e \in E(c, x)$ is a declarative sentence or question that is interpretable in context $\bar{c}$; (2) each $e \in E(c, x)$ has the same truth-conditional meaning in $\bar{c}$ as the meaning of sentence x in context c.

Note that the sentence x will most often in this article be a system sentence $s_{i,j}$ in linguistic context $c_{i,j}$, but can also be a user sentence $u_{i,j}$ in linguistic context $c'_{i,j}$.

Thus, each $e \in E(c, x)$ is a paraphrase of x that is interpretable in the linguistically empty context and that preserves the truth-conditional meaning of x in context c. Note that $E(c, x)$ is a set because there may be multiple ways of paraphrasing x that are equivalent in meaning. Given an equivalence relation between sentences that identifies

4 An equally plausible definition would be to define $c_{i,j}$ to also include the following sentences within utterance s_i, that is, $s_{i,j-1}, s_{i,j+1} \ldots s_{i,|s_i|}$ (and an analogous definition for $c'_{i,j}$). That is, the context would be extended to include sentences that follow $s_{i,j}$ in the utterance s_i. This would allow instances of cataphora, for example, to be handled in the definitions of explicatures and attribution.

whether any two sentences are equal in meaning or not, we can think of a single member of $E(c, x)$ as a representative of the entire set $E(c, x)$. Following this, in a slight abuse of terminology we will henceforth often write "the explicature of x in context c is e" as if there is a single unique explicature e, with the understanding that e represents the entire set $E(c, x)$. We will also write $E(c, x) = e$ as shorthand for $E(c, x)$ being equal to the set of all sentences whose meaning is the same as that of e.

In addition, we define interpretability as follows:

Definition 7 (Interpretability)
A sentence x in context c is **uninterpretable** if the truth-conditional meaning of x in context c is unclear. In this case we write $E(c, x) = \texttt{NULL}$.

Figure 1 shows several examples illustrating these definitions. Some key points are as follows:

Remark 1: In example E1, the system response is a direct answer to a question, $s_{2,1} = 25$ *years old*. $s_{2,1}$ itself is not a declarative sentence, but given the context (in particular the question it is answering), its explicature is the standalone proposition *George Harrison was 25 years old when "Wonderwall Music" was released*. This type of example—where a direct answer to a question is an entity, noun-phrase, or some other fragment, but its explicature is a standalone proposition—is important and frequent. As another example consider the following:

u_1**:** What was George Harrison's first solo album?
s_1**:** Wonderwall Music
$E(c_{1,1}, s_{1,1}) =$ *George Harrison's first solo album em was "Wonderwall Music"*

Remark 2: In Example E3, the system segment is a sequence of two declarative sentences. Each sentence has an explicature that is a standalone proposition. This type of case is again frequent and important.

Remark 3: In Example E4 the system utterance is uninterpretable, because it is not clear what "the band" is referring to. Example E5 contains disfluencies that make it difficult to reliably interpret: "it" is not the expected pronominal reference; in this context "25" becomes too ambiguous to interpret as referring to the age of a human entity.

Remark 4: Examples E6 and E7 contain questions in the system and user utterance, respectively. These examples illustrate that single questions (E7) or questions within multi-sentence utterances (E6) have well-defined explicatures.

3.3.1 The Full Definition of AIS. With this background, we can now give the full definition of AIS:

Definition 8 (AIS, full definition)
A pair (s, c), where s is a sentence and $c = (c_l, t)$ is a pair consisting of a linguistic context and a time, is Attributable to Identified Sources (AIS) iff the following conditions hold:

1. The system provides a set of parts P of some underlying corpus K, along with s.

2. s in the context c is interpretable (i.e., $E(c, s) \neq \texttt{NULL}$).
3. The explicature $E(c, s)$ is a standalone proposition.
4. The pair $(E(c, s), t)$ is attributable to P.

The pair $(E(c, s), t)$ is **attributable** to a set of parts P of some underlying corpus K iff: A generic hearer will, with a chosen level of confidence, affirm the following statement: "According to P, $E(c, s)$", where $E(c, s)$ is interpreted relative to time t.

The definition is very similar to the earlier definition of AIS for standalone propositions, but with checks for interpretability, and with attribution applied to explicatures of system sentences. Note that AIS can only hold for system sentences that have an explicature that is a standalone proposition (condition 3). For example, the explicature in Example E6 in Figure 1 is not a standalone proposition, as it is a question. We leave the treatment of cases such as these to future work (we might for example evaluate attribution for declarative sentences within the explicature, excluding questions; or we might evaluate presuppositions within the questions themselves).

3.3.2 Attribution of Entire Utterances. In the previous sections we have described AIS for the individual sentences $s_{i,1} \ldots s_{i,|s_i|}$ within a system utterance s_i. This assumes that such a segmentation of the utterance into sentences is available, for example, it is provided by the system. An alternative is to evaluate entire utterances s_i for AIS, in a "single-shot" annotation. AIS applied at the utterance level could potentially have the advantages of simplicity, and the avoidance of segmenting utterances into sentence boundaries. It has the potential disadvantage of being coarser grained, not allowing AIS judgments at the sentence level. The choice of sentence-level vs. utterance-level AIS will depend on the exact application of AIS.

It should be relatively straightforward to extend the full definition of AIS (Section 3.3.1) to apply to multi-sentence utterances. The definition of explicatures would need to be extended to multi-sentence utterances; the definition of standalone propositions would also have to be extended to apply to multiple sentences; the definition of "attributable" would also need to be extended.

3.4 Operationalization of AIS

In the above definition of AIS, three definitions are of key importance: (1) the "according to" test for standalone propositions; (2) the definition of interpretability; (3) the definition of explicatures, which are related to the interpretation of utterances in non-empty linguistic contexts.

Note that it is not necessary for annotators to explicitly wield all of these definitions, or come to understand any of them in entirely formal terms, in order to provide AIS judgments. Instead, we create annotator guidelines that convey these concepts in an intuitive way that is accessible to annotators who may have limited or no linguistic background. In Section 4.1, we provide a detailed overview of the task design, and reproduce the full annotation guidelines in the Appendix. Here, we highlight some important connections between the formal definitions and the guidelines:

- Figure 2 shows how the overall task is framed to the raters, as a two step process of (A) determining "the information provided by the system

Definition. Attribution of a system-generated response in relation to the source document can be established by considering the following:

A. What is the **information provided by the system response?**

B. Is this information an **accurate representation of information in the source document?**

Figure 2
An excerpt from the annotation guidelines, showing the two steps in the task. Step A corresponds to ascertaining the explicature of the system response; Step B corresponds to the "according to" test.

response", and (B) determining whether this information is "an accurate representation of information in the source document". Step A corresponds directly to determining the explicature of the system response, and Step B corresponds directly to an application of the "according to" test.

- The guidelines go into some detail in describing what is meant by "the information provided by the system response". In particular, several examples are given in the guidelines showing "paraphrase(s) of the information provided by the system response", which is essentially identical to the notion of explicature. See Figure 3 for some of these examples.
- Finally, the guidelines give a description of the "according to" test that is very close to the formal definition. See Figure 4 for an excerpt of this text.

We also note that the guidelines give guidance to raters regarding the decision of whether or not a system response is interpretable (Section 4.1.1).

4. Human Evaluation Study

We evaluate the feasibility of human AIS assessment for three NLG tasks: conversational question answering, summarization, and table-to-text generation. To examine if human judgments are stable enough to distinguish different technical models with statistical reliability, we present evaluators with the output of several models for each of the tasks.

The set-up for these annotation tasks is to ask annotators to rate the AIS quality of s, some model produced output given some attributed source P. In the conversational QA and summarization settings, P is a document or passage from a document, while in the table-to-text setting P is a table and its description. For conversational QA, annotators are also provided with a context c, which is the set of previous conversation turns. c is used to help annotators understand the contextualized meaning of the model output, what we formally define as explicature in §3.3.

User query 1
which network is days of our lives on
System response 1
the show is on NBC
Paraphrase of information provided by system response
days of our lives is on NBC

User query 2
which network is days of our lives on
System response 2
NBC
Paraphrase of information provided by system response
days of our lives is on NBC

User query 3
how many seasons are there of days of our lives
System response 3
there are 56 seasons
Paraphrase of information provided by system response
there are 56 seasons of days of our lives

Figure 3
Some examples of system responses shown to annotators, paired with paraphrases of "information provided by the system response". This corresponds directly to the explicature of the system response.

Because this is a challenging task with many possible edge cases (such as those discussed in Table 1), we ask five annotators to judge each example. In our results section, we compare to the consensus answer (if there is one) for simplicity. In future work, researchers who wish to use AIS for evaluating systems might find it more economical to distinguish cases that are more clear-cut (i.e., unanimous) and those where there may be some inherent ambiguity.

4.1 Task Design

We break the annotation task into two stages described in §4.1.1 and §4.1.2, which mirrors the formal steps in the AIS definition (§3.3.1) First, the annotators are asked if they are able to understand and identify the information being shared in the model output without seeing the source document (i.e., whether it is *interpretable* on its own). Then, if the output is deemed interpretable, the annotators are shown the "identified source" P and asked whether all of the information that is shared in S can be attributed to P (i.e., whether it is *AIS*). As described in the results sections, the splitting of the task into these two steps helps annotators to first filter out outputs that are badly formed (e.g., ungrammatical to the point of impeded intelligibility) or too ambiguous (e.g., unclear pronouns) to appropriately evaluate the attribution. In the results, we report scores based on the annotator consensus (i.e., majority vote): the percent of total examples marked as interpretable (*Int* in Tables) and the percent of interpretable examples that were marked as AIS (*AIS*). In some datasets, certain examples were flagged as difficult to annotate due to legibility-related issues (see §4.1.3). For those cases, we separately

B. Definition of "An Accurate representation of Information in the Source Document". Again, you should use your best judgment in determining whether all of the information provided by the system response is "an accurate representation of information in the source document." We give the following guidance:

- In determining this question, ask yourself whether it is accurate to say "the document says…" or "according to the document…" with the system response following this phrase. For example, is it accurate to say "according to the document below, In the American daytime drama Days of Our Lives, Doug Williams and Julie Williams are portrayed by Bill Hayes and Susan Seaforth Hayes" in the example given above?
- Be sure to check **all** of the information in the response. If only some of the information is supported in the document, but other parts of the information are missing from the document or not an accurate representation, then please mark "**No, not fully attributable**."
- The concept of "accurate representation" should be close to a journalist's conception of this phrase. For example take this excerpt from this page on Accuracy in the NPR Ethics Handbook: "When quoting or paraphrasing anyone…consider whether the source would agree with the interpretation…" In other words, if you had written the source document, consider whether you would view the system response as an accurate representation of information in that source document.

Figure 4
Instructions given to the raters for Step B, corresponding to the "according to" test.

report the percentage of examples that were flagged (*Flag*) and thus excluded from the interpretability and AIS scores.

4.1.1 Interpretability Rating. In the initial stage of the annotation task, we show the annotators the model output s and any preceding context c without showing the source. We ask them to identify the interpretability by posing a yes/no question. For example, in the summarization task the annotators are asked:

*Is **all** of the information relayed by the system summary interpretable to you?*

Note that context c is populated with preceding turns of the system–user interaction[5] in the conversational QA task, whereas in summarization and table-to-text tasks it is always empty. In the instructions for the conversational QA task, we give them additional instructions that explicitly call out how to use the context c in interpreting output s.

Here, the goal is to tease out if the model-generated output s contains any potential ambiguity that would prevent or misguide establishing attribution to its source P. Anaphora resolution is the main source of this type of ambiguity, where deictic elements do not have clear antecedents within s or its context—for example, pronominal usage with an unclear or broken coreference chain or definite noun phrases as first

5 Some interactions may contain no previous turns.

mentions. Additionally, syntactic ambiguity or disfluency may also result in diminished interpretability of *s* (see Examples E4, E5 in Figure 1).

We acknowledge potential anthropomorphizing effects on how annotators interpret the system output (Gopnik and Wellman 1992). Because cooperative meaning co-construction between interlocutors is the default communicative strategy of inter-human interaction (Grice 1975), when faced with ambiguities and slight discrepancies in the system output, annotators may be "forgiving" of diminished interpretability, especially if the underlying source is present and can help recover missing context.

In our experiments we have found that not presenting the attributed source documents (P) at this stage is crucial for ensuring that evaluators are strict in their assessment of interpretability of the system output (see Figures E.6, F.8, and G.10 for how it was implemented in the task interface).

4.1.2 AIS Rating. If an annotator selects "yes" for the interpretability question, we show them the source *P* and ask them whether **all** of the information relayed in the output *s* can be supported by *P*. For example, in the conversational QA task the annotators are asked:

*Is **all** of the information provided by the system response fully supported by the source document?*

Note that the *P* for the conversational QA task is the retrieved document that serves as the source of the system output *s*. In the summarization task *P* is the original news article from which the summary in *s* was derived. In the table-to-text task *P* is the original table, highlighted cells, and table metadata (table title, section title, and section text) from which the textual table description is generated.

In the instructions, we tell annotators to first think about all of the information that is contained in the output including: what's directly stated in the output sentence verbatim as well as any explicatures that can be made from the output with respect to the context, such as inferring pronoun references from the conversational history.

Annotators are instructed to only mark output as attributable if it is clear that all parts can be directly inferred from the source. The instructions specifically describe the same "according to" test that we used in our definitions in Section 3:

In determining this question, ask yourself whether it is accurate to say "the provided news article says…" or "according to the news article…" with the system summary following this phrase.

If the output is misrepresenting information from the source because it is misleadingly worded, missing important context, or even changing only slight details, these cases are all counted as "not fully attributable".

4.1.3 Flag Rating. A special rating is reserved for flagging items that would be disqualified from the task altogether because they flout the range of possible relationships between the utterance, its context, and the source defined in Section 3.3.1.

In practical terms, these are tasks that are too malformed for annotators to perform judgments on. This category includes tasks with rendering issues in the interface (missing task elements, e.g., empty utterance), corrupted text resulting in non-communicative utterances (bad text encoding, HTML artifacts), underspecified source (the source document itself is ambiguous because it is too short and may contain unresolved reference chains), or a source that is difficult to understand because it requires expert-level knowledge.

Once a task is flagged, it is disqualified from the rating queue of the annotator who flagged it. Other annotators may choose not to flag this item; cumulative ratings and

interannotator agreements are calculated for all non-flagged ratings of a task (see the flag sections in the annotation guidelines in the Appendix).

4.1.4 Limitations. By asking yes/no questions, we can greatly reduce the complexity of this task for annotators. However, for some applications of AIS measures, it may be useful to have more fine-grained measures. Additionally, we ask annotators to evaluate the entire output (rather than sentences or specific spans) under the reasoning that if even one span within the model output is not AIS, then the whole output is not AIS (cf. Maynez et al. 2020; Durmus, He, and Diab 2020).

We also acknowledge that there are other aspects of model output quality (e.g., relevance, non-redundancy) not evaluated here. We focus on the separate evaluation of AIS as part of a focused effort toward quantifying the attribution itself, disentangled from other desirable generation qualities.

4.2 Human Evaluation Procedure

The ratings were performed by a group of nine paid full-time annotators[6] under the guidance and supervision of a project manager. The annotator team is based in Hyderabad, India; the annotators are native speakers of the Indian dialect of the English language. The annotators do not have a background in linguistics. They were trained for this specific task.

Three separate user interfaces were developed for performing the evaluation in this study: one for the conversational QA tasks evaluating the output of models trained on QReCC and WoW datasets, another for summarization tasks evaluating the output of models trained on the CNN/DailyMail (DM) dataset, and lastly one for table-to-text tasks evaluating the output of models trained on the ToTTo dataset. The interfaces share many design elements with task-dependent modifications. For example, the conversation QA interface shows the conversational history. All three interfaces explicitly hide the source document/table at the stage when interpretability of the system output is evaluated (see the Appendix for the interface layouts and annotator prompts in Figures E.6, E.7, F.8, F.9, G.10, and G.11).

The annotators were trained on the tasks in a series of stages. First, a pilot study of 50–100 items was conducted with the first iteration of the annotator instructions. As part of the pilot, all ratings were required to have written justifications elaborating the reasoning for the provided rating. The results of the pilot were analyzed by the authors to identify common error patterns; the collected justifications were helpful in understanding the reasoning annotators used to arrive at their ratings. The results of the review were communicated back to the annotators, and the instructions were modified to emphasize areas leading to common ratings errors.

Next, a portion of the ratings was inspected by the authors for persistent error patterns and the feedback communicated to annotators. Additionally, the annotators collected edge cases where they found it difficult to make judgments. These edge cases were adjudicated by the authors; recurring complex patterns were used to expand the annotation guidelines (see the Appendix for the full final instructions).

6 The annotators are not paid by Google directly, but by Google's suppliers. These suppliers are responsible for managing the annotators and setting their compensation, and are obligated to comply with Google's Supplier Code of Conduct (`https://about.google/supplier-code-of-conduct/`).

Finally, the annotator team performed internal audits on a subset of completed tasks.

Annotators were initially trained on the conversational QA tasks; other tasks and training were introduced subsequently.

5. Experiments

In the following section, we demonstrate the utility of AIS by showing how it can be applied to three different tasks (conversational QA, summarization, and table-to-text generation) in which the model output is—by design—always meant to be attributable to some source document. We instantiated the AIS annotation task for four datasets in these domains (see Table 2) and asked human annotators to evaluate the generated outputs from multiple models. In order to show the applicability of AIS in detecting nuanced differences between different types of model outputs, we specifically chose models for each dataset that would represent a range of different types of outputs rather than just selecting a set of state-of-the-art models. We also annotated a selection of gold references from each dataset to better understand the AIS quality of existing datasets in these areas. We end with analysis of how effectively humans can annotate AIS as well as a discussion of various interpretability and AIS patterns that we found in the resulting annotations.

5.1 QReCC Answer Generation

Setup. We use the QReCC dataset (Anantha et al. 2021), a collection of multi-turn conversational QA interactions that extends conversations coming from NaturalQuestions (Kwiatkowski et al. 2019), QUAC (Choi et al. 2018), and CAST-19 (Dalton et al. 2020). In this task, a model is given a conversational history and generates a contextualized response. We use a task setup where the document passage containing the answer to the current query has already been retrieved (using the oracle retrieved document passage as the attributed source). We use different variations of T5 models including both base

Table 2
Summary of tasks used in human annotation study.

Task	Dataset	Context (C)	Source Document (P)	System Output (S)
Conversational QA	QReCC (Anantha et al. 2021)	Conversational History	Retrieved Document	Response
Conversational QA	Wizard of Wikipedia (Dinan et al. 2019)	Conversational History	Retrieved Fact	Response
Summarization	CNN/DM (Nallapati et al. 2016)	N/A	Source Article	Summary
Table-To-Text	ToTTo (Parikh et al. 2020)	N/A	Table, Table Description	Caption

Table 3
Results of human study on 200 examples from QReCC test set (randomly sampled from set of examples where conversation length ≤ 5 turns). PT= pretrained model, FT = fine-tuned on QReCC training data.

Model	Size	Int	AIS
T5-PT (with Evidence)	Small	43.0*	82.6
	Base	47.0*	69.1*
T5-FT (no Evidence)	Small	57.8*	25.2*
	Base	59.8*	21.8*
T5-FT (with Evidence)	Small	99.0	87.9
	Base	98.0	87.2
Reference		99.0	87.8

*Indicates that the result is significantly lower than the **highest score** in the column (with $p < 0.01$).

and small size variants. First, we use the pre-trained T5 models (PT) by themselves by prompting the model (formatted as: "Query:... Conversation History: ... Document: ... Answer:"). We also use a version of T5 that has been fine-tuned on QReCC (FT) which uses special tokens to separate the query, context, and document instead of natural-language prompts. Lastly, to sanity-check the AIS measures, we use a version of the model (no evidence) that only sees the query and conversation history but not the document at generation time. We expect that the AIS subscores should be much lower in the model that does not use the evidence from the document to generate the answer.

Results. We show results in Table 3. The model outputs' interpretability increases substantially after fine-tuning (by about 50 points). The AIS subscore is highest in the fine-tuned model that uses evidence in its input. As expected, AIS is drastically lower in the model that does not use the document as input at generation time (the no-evidence model) which is both interpretable and AIS only 15% of the time. Differences between model sizes (small vs. base) are generally not significant except for the pretrained-only model, though the AIS scores of the smaller versions are typically slightly higher.

5.2 WoW Answer Generation

Setup. We used the seen portion of the test set from Wizard of Wikipedia (Dinan et al. 2019). In this task, a model is given a conversational history and generates a contextualized response based on information from Wikipedia. As with QReCC, we again use a setup where the Wikipedia sentence has already been retrieved (using the oracle retrieved sentence as the attributed source). To avoid chit-chat style utterances that may not be sharing new information, we sampled 200 examples per model where the previous utterance was a question (contains '?'). We used the models from Rashkin et al. (2021). That paper introduced a controlled T5 model trained on the Wizard of Wikipedia data which uses control tags and re-sampling to target model output that is more faithful to the document (by looking at heuristics such as entailment metrics, lexical precision, and first-person usage). Similar to that paper, we also compared with three models that are seq2seq-style conversation models: the original answer generation system from Dinan et al. (2019), the dodecaDialogue multitask system from Shuster et al. (2020), and a T5-base model (Raffel et al. 2020) fine-tuned on Wizard of

Table 4
Results of human study on 200 examples from Wizard of Wikipedia test set (Dinan et al. 2019) (the seen topic split, using only conversation turns where the previous turn has a question mark).

Model	Flag	Int	AIS
WoW Baseline (Dinan et al. 2019)	4.0	84.4*	19.8*
Dodeca (Shuster et al. 2020)	8.5	100.0	60.1*
T5 (Raffel et al. 2020)	5.5	98.4	39.8*
T5 (with Controls) (Rashkin et al. 2021)	7.5	99.5	92.4
Reference	4.0	100.0	15.6*

* Indicates that the result is significantly lower than the **highest score** in the column (with $p < 0.01$).

Wikipedia data. Because the model from Rashkin et al. (2021) was specifically trained to be more faithful to evidence, we expect that it will score higher in the AIS category.

Results. We show results in Table 4. Compared with the QReCC data (in which only a few examples were flagged), more examples were flagged with the Wizard of Wikipedia data, which we included as an extra column. The general trend of results is similar to what was found in the human evaluations of faithfulness and subjectivity in Rashkin et al. (2021). As expected, the model that has specific controllable inputs for increasing the model's faithfulness to the input document achieves the highest the AIS scores overall. We also note that the AIS scores of the gold references is lower than the model outputs. We discuss this more in Section 5.5.5.

5.3 CNN/DM Summarization

Setup. We extend our evaluation framework for a second task, summarization, to confirm that AIS can be more broadly applicable. AIS is crucial in summarization where a generated summary (S) must be well-supported by the source article (P). In contrast to some of the prior work in hallucination evaluation in summarization (Durmus, He, and Diab 2020; Maynez et al. 2020), the annotators in our task evaluate the full summary for attribution (rather than at a sentence-level or a span-level), in order to account for cases where two individual text spans may be attributable to a source document but—when composed together—convey information that is different from the source document (misordered events, pronouns that no longer have the correct references when misordered, etc.). As a first step in applying AIS to summarization, we compare the performance of three different approaches (abstractive vs. extractive vs. hybrid) on 200 examples randomly sampled from the CNN/DM (Nallapati et al. 2016) test set. The source articles in this dataset come from articles in CNN and DailyMail news and the summaries are extracted from bulleted highlights that were included with the article by the journalists. We expect that high-quality AIS annotations will show a trend where extractive systems achieve higher AIS scores because they are copying directly from the source without adding anything. First, we used MatchSum (Zhong et al. 2020), a state-of-the-art extractive summarization model. Because this model is extractive, it is expected that it will be the least prone to hallucinations. We also used an abstractive summarization system, BigBird (Zaheer et al. 2020). Lastly, we used Pointer-generator Networks from See, Liu, and Manning (2017)—a hybrid approach that uses

Table 5
Results of human study on 200 examples from CNN/DM test set (randomly sampled). Of the three models we tested with, unsurprisingly the more extractive models have higher AIS scores.

Model	Approach	Int	AIS
MatchSum (Zhong et al. 2020)	Extractive	90.0	99.4
Pointer-Gen (See, Liu, and Manning 2017)	Hybrid	90.0	97.8
BigBird (Zaheer et al. 2020)	Abstractive	90.0	87.2*
Reference	–	86.0	54.1*

* Indicates that the result is significantly lower than the **highest score** in the column (with $p < 0.01$).

an abstractive seq2seq model but with an explicit copy mechanism that can extract information from the source document.

Results. We show results in Table 5. The more extractive approaches generally reach higher AIS subscores. This is a somewhat expected result—extractive systems are less likely to output hallucinations as they are quoting information verbatim from the documents. As with Wizard of Wikipedia, the AIS scores of the gold reference summaries is surprisingly lower than the model output, which we will discuss more in Section 5.5.5.

5.4 Table-to-Text ToTTo data

Setup. Lastly, we show the utility of extending AIS to a table-to-text task where P is a table rather than a text document. S is a sentence generated by a model to describe some highlighted portion of the table. We chose the ToTTo dataset (Parikh et al. 2020), testing with T5 and ByT5 models that were previously used with this data in the GEM benchmark (Gehrmann et al. 2021). We experiment with two different sizes of ByT5 and three different sizes of the T5 architecture. As before, we sampled the output of 200 examples from the test set. We also annotated 200 ground-truth references from examples in the dev set (as the test set does not have gold-truth references publicly available).

Results. We show results in Table 6. The model with the most "interpretable" responses was T5-base, with the ByT5 architectures being significantly less interpretable. On the other hand, the T5 architecture responses were more likely to be flagged (according to the annotators this was because the flagged responses contained artifacts like unintelligible character encoding errors). Generally, we do not observe statistically significant differences in the AIS subscores though the larger architectures tended to have slightly lower AIS scores (similar to our observations of Table 3).

5.5 Annotation Quality

In this section we discuss the further implications of the human annotation results. We focus on two primary questions: (1) can humans reliably annotate AIS? and (2) what do our measured AIS ratings indicate about NLP data and models?

Table 6
Results of human study on 200 examples from ToTTo test set (model output) and development set (ground-truth references).

Model	Flag	Int	AIS
ByT5-Base	0.0	78.9*	88.5
ByT5-XL	0.0	79.5*	86.2
T5-Small	3.0	86.5	88.6
T5-Base	5.0	91.1	86.6
T5-XL	6.0	89.4	85.1*
Reference	0.0	83.9	91.0

* Indicates that the result is significantly lower than the **highest score** in the column (with $p < 0.01$)

5.5.1 Interannotator Agreement. We show the interannotator agreement (IAA) for crowd annotators in the left half of Table 7. The metrics we used include Krippendorff's alpha comparing individual ratings, pairwise agreement (PA) comparing individual ratings, and an F1 score comparing individual ratings to the consensus (majority vote). Agreement is generally moderate to high, displaying that—while this is a challenging task—the annotators are able to be fairly consistent with one another. The alpha scores are generally lowest on the summarization CNN/DM task, perhaps because the output text is much longer in summarization, increasing the complexity of the rating task. The F1 scores are similarly high, particularly on the AIS ratings.

5.5.2 Audits. Separately, the annotator team also performed internal audits on the annotation quality where a project lead from the annotator team examined a sample of individual annotator judgments at different points (snapshots) of the annotation process (Table 8). QReCC and WoW annotations were evaluated together as the broader conversational QA annotation task. The overall reported quality is in the high nineties for all three tasks with slight variations. The annotation quality for the conversational QA tasks remains high across all snapshots; we attribute this to the annotators' extended

Table 7
Annotator agreement measured as interannotator agreement (left half of the table) or as agreement with expert consensus (right half of the table, only measured on QRECC and CNN/DM tasks). Metrics include—F1: a F1 measure comparing individual ratings to the consensus rating; PA: pairwise agreement as percentage of individual pairs that agree; α: Krippendorff's alpha measure comparing pairs of individual ratings.

	IAA						vs. Expert					
	Int			AIS			Int			AIS		
Task	F1	PA	α	F1	PA	α	F1	PA	α	F1	PA	α
CNN/DM	.83	.80	.46	.92	.89	.69	.48	.60	−.04	.81	.86	.61
QReCC	.97	.96	.91	.93	.89	.76	.77	.81	.54	.77	.78	.54
WoW	.88	.93	.60	.95	.88	.79	–	–	–	–	–	–
ToTTo	.95	.95	.84	.92	.92	.74	–	–	–	–	–	–

Table 8
Quality measure on samples of annotations for conversational QA, summarization, and table-to-text tasks. Snapshots (*S*) represent consecutive annotation sprints with individual annotator judgments (*n*) replicated at 5 per task. A sample (*Sample*) of each snapshot was evaluated by a project lead on the annotator team. The quality of annotations (*Quality*) was assessed over a varying number of snapshots for each task. The evaluated annotations exclude flagged tasks.

	Conversational QA			Summarization			Table-to-Text		
S	*n*	Sample	Quality	*n*	Sample	Quality	*n*	Sample	Quality
1	642	.04	1.00	88	.06	.67	261	.08	1.00
2	726	.05	1.00	339	.06	.87	2,518	.19	.96
3	1,895	.03	1.00	469	.10	.94	2,463	.30	.97
4	2,520	.04	.97	682	.08	1.00	1,151	.34	.96
5	–	–	–	608	.08	1.00	849	.30	.94
6	–	–	–	652	.08	1.00	–	–	–
7	–	–	–	928	.03	1.00	–	–	–
Total	5,783	.04	.99	3,766	.07	.98	7,242	.26	.96

experience with the task prior to the annotation of this dataset.[7] The quality of the summarization annotations shows an increase over snapshots, as annotators internalize the guidelines and gain expertise in the task. The quality of the table-to-text annotations fluctuates and is generally the lowest of the three tasks; we attribute this to a much larger sample for which quality was measured. Overall, across the three tasks, the larger the evaluated sample, the lower the overall reported quality. Barring genuine task differences that would lead to variations in annotation quality, this suggests that the reported table-to-text quality of annotations is the most representative of all three tasks.

5.5.3 Annotator Performance. Average task completion times decreased across all types of tasks as the annotators were exposed to more tasks and internalized the instructions (Table 9). Only the initial pilots for the three tasks included time-consuming, required justifications.

At the same time, the absolute task completion times are consistently and substantially different across the three tasks, suggesting their uneven complexity, with conversational QA taking the shortest amount of time to complete, summarization requiring the longest, and table-to-text falling in-between. This pattern follows the trend in the distribution of inter-annotator agreement across the three tasks: Tasks with shorter completion times generally have higher interannotator agreement. We postulate that this is primarily due to the difference in the amount of context that is necessary to perform ratings. Although conversational QA tasks may contain several turns of preceding interactions between the system and the user as well as the source document, the amount of information in the source articles in the summarization task is substantially larger. Likewise, source tables in the table-to-text task can be extensive and have the added information complexity of cell highlighting and table metadata. Finally, register and discourse structure effects may be at play here as well. Conversational QA

7 The annotator pool was involved in annotating a series of related tasks for Conversational QA beyond the reported results in this article.

Table 9
Average completion times (*ACT*) for rating tasks in seconds. Conversational QA tasks include evaluation of generated text for QReCC and WoW. Summarization tasks include evaluation of generated text for CNN/DM. Table-to-text tasks include evaluation of generated text for ToTTo. Justifications were required for all question tasks at the pilot stage, but not at the production stages. Average completion times decrease for all three task types as annotators gain more experiences over the amount of observed tasks (*Tasks*), but are always relatively longer for summarization. Note that the average completion times may be reduced further for summarization with more tasks observed by annotators, a pattern we see in the conversational QA and table-to-text task types.

		Conversational QA		Summarization		Table-to-Text	
Stage	Justification	Tasks	ACT, sec	Tasks	ACT, sec	Tasks	ACT, sec
Pilot	+	75	375.30	50	687.02	100	324.72
Start	−	762	136.76	187	308.14	136	238.40
Finish	−	4,022	73.19	900	263.55	1,496	193.95

tasks build upon colloquial interactions between the user and the system, setting up the context of the interaction in shorter utterances and helping annotators anticipate the contents of the source document. Likewise, Wikipedia, news articles, and tables package information differently as they serve somewhat different communicative goals, and it is possible that one of these source types is more amenable to inspection required for performing AIS ratings.

5.5.4 Expert Ratings. Where AIS is used as a metric for ranking generative models, the internal consistency of crowd annotations is paramount. But, to help illuminate the inherent challenges in calibrating this annotation task, we also compare the crowd ratings with those of experts on a small set of examples. Due to the challenges of scaling expert evaluations, we limited expert ratings to two tasks (CNN/DM and QReCC) with 50 examples each. The experts (two co-authors) first annotated the examples separately from each other using the same interface as the crowd annotators and then discussed their answers to reach a consensus. *Expertise* here might be derived from general educational background (a different approach to close reading), the ability to discuss annotations (and to do so carefully at self-guided pace), specialized knowledge, and first-hand familiarity with the evaluation framework. Expertise does not imply that the experts have more experience performing the task than the crowd annotators.

In order to account for natural ambiguity in assigning a rating category, experts marked some cases as "either option acceptable". We compare the individual crowd annotator ratings to the expert consensus in the right half of Table 7. Crowd annotators tend to agree with each other more than they agree with experts, which is expected due to differences in background, incentives, and procedure, although there is still reasonably consistent agreement in most cases. On closer inspection, we find that most disagreements are cases where there is underlying ambiguity caused by vagueness in the evidence or model output. In these cases experts erred more on the side of being *critical* of the model and crowd annotators erred more toward being *lenient*. In the case of conversational QA, most of the AIS disagreements involved cases where the document and the response do not refer to an entity using the same naming conventions (e.g., using both first and last name; see Table 10) leaving some ambiguity that the document is

Table 10
Selected conversational QA examples marked as **AIS** by annotators but **not AIS** by experts, where differences in how entities are named (first name vs. last name) in the response vs. document may cause some ambiguity as to whether the document is accurately describing the same person. Some people may also have background knowledge about the entities involved that helps them disambiguate whether these documents are referring to the same "mayer" or "adams" as the question.

Example 1
Query
what is a charity john mayer works for?
Response
in 2002, john mayer created the back to you fund, a nonprofit organization that focuses on fundraising in the areas of health care, education, the arts and talent development.
Attributed Document
personal life on his third episode of current mood , mayer revealed he had been sober for two years. he stopped drinking after having what he says was a six-day hangover after drake 's 30th birthday party. relationships mayer has never married, but has been romantically involved with jennifer love hewitt , jessica simpson , minka kelly , jennifer aniston , and katy perry . philanthropy in 2002, mayer created the "back to you" fund, a nonprofit organization that focuses on fundraising in the areas of health care, education, the arts and talent development. the foundation auctions exclusive john mayer items, such as guitar picks, t-shirts and signed cds. . . .

Example 2
Conversation History
USER: where did john adams practice law?
SYSTEM: harvard
Query
what happened after john adams practiced law at harvard?
Response
in 1756, john adams began reading law under james putnam, a leading lawyer in worcester.
Attributed Document
law practice and marriage abigail smith adams – 1766 portrait by benjamin blyth john adams – 1766 portrait also by blyth in 1756, adams began reading law under james putnam, a leading lawyer in worcester. in 1758, he earned an a.m. from harvard, and was soon admitted to the bar, having completed his studies under putnam. he developed an early habit of writing about events and impressions of men in his diary; this included james otis jr. 's 1761 legal argument challenging the legality of british writs of assistance , allowing the british to search a home without notice or reason. . . .

referring to the same entity as the response. The greatest source of disagreements overall is the interpretability question in the summarization task (see examples in Table 11). The summaries in the CNN/DM dataset were originally crawled from high-level article highlights, and experts observed that—due to the linguistic style of these highlights—there were many cases where the language may be vague or ambiguous, making this dimension more challenging. Because we use interpretability as a pre-filtering stage for the AIS question, we make allowances for the annotators being more inclusive. Despite the differences on the interpretability dimension, they generally agreed with experts on most AIS questions, our primary evaluation dimension.

5.5.5 Limitations of Gold References. The last rows in Tables 3, 4, 5, and 6 show annotation results on reference answers sampled from these datasets. The results demonstrate that there is actually a limit on the AIS quality of the data itself in multiple tasks. We include

Table 11
Selected summary examples marked as interpretable by annotators but non-interpretable by experts. We note that the style of language in these summaries can be vague which may increase the difficulty in leaving a binary interpretability judgment.

Example 1
Summary
deciding who you will vote for may have more to do with your family than who won the leaders debate (above) finds study which looked at the voting habits of twins born in the uk . the aim was to explore how much nature and nurture influence our party political allegiances and potential voting preferences
Example 2
Summary
Charlie Stayt was broadcasting live from a primary school in Southampton . He missed out the letter 'c' when he scrawled the word on a whiteboard . Outraged viewers took to Twitter to complain about the spelling error . Stayt later described the gaffe as 'one of those things'
Example 3
Summary
university lecturer dr alex russell shares his expert advice . dr russell says that anyone can improve their tasting skills in four hours .
Example 4
Summary
in fact , it 's an advert from cosmetics giant revlon for their latest lipstick . the stylish ad is filmed entirely in black and white , with just a slick of pink visible on the woman's lips. revlon uk 's new global tag line , love is on , is the label 's first major relaunch in more than a decade .

examples of non-AIS references in Table 12 to illustrate what some of these examples look like. We hypothesize that this is because the originators of the data were not specifically instructed to be as faithful to the underlying documents as possible. In the case of Wizard of Wikipedia (Dinan et al. 2019), the gold response was fully AIS in 16% of the cases we measured. But, this may be because that dataset was constructed with slightly different objectives—to contain *both* informative and engaging responses. The crowdworkers who created the data were provided documents to enhance their conversations but could do so at their own discretion, often including their own thoughts and experiences in the conversation as well. This is also reflected in the CNN/DM AIS scores—summaries in CNN/DM are only attributable to the documents in 54% of the interpretable examples. Looking more closely, we speculate that this may be due to the post-hoc data creation process used to extract summaries from article highlights written by journalists. We observed that the reference summaries in CNN/DM may sometimes refer to external pieces of information that may have accompanied the article (a picture, a headline, etc.) or sometimes make assumptions about what the intended audience of the article might already know that can affect either the interpretability or AIS scores (see Example 1 in Table 12 and Example 3 in Table A.17). These results indicate that there is still a need for high-quality AIS data for training new NLG models.

Table 12
Examples of **reference** (gold) text that was marked by annotators as **interpretable but not AIS**.

Example 1. (CNN/DM)

Summary *(Reference)*

thomas piermayr has been training with blackpool this week .
austrian defender is a free agent after leaving mls side colorado rapids .
blackpool are bottom of the championship and look set to be relegated . .

Attributed Document

Blackpool are in talks to sign Austria defender Thomas Piermayr.
The 25-year-old has been training with the Championship club this week and they are keen to get him on board for what is expected to be confirmed as a campaign in League One next season.
Piermayr is a free agent and had been playing for Colorado Rapids.
The former Austria U21 international had a spell with Inverness Caledonian Thistle in 2011.
Thomas Piermayr (left, in action for the Colorado Rapids) tries to tackle Obafemi Martins last year . .

Explanation: The article doesn't mention Blackpool being the bottom of the championship but it is included in the gold summary. Possibly the journalist who wrote the article highlights (that the summary data was extracted from) included that sentence because they were assuming that the sports section audience might already be aware of the standings.

Example 2. (Wizard of Wikipedia)

Query

i really love kentucky, i was born and raised here, have you ever been there?

Response *(Reference)*

my parents are from kentucky! i have only been to east, south-central region state a few times.

Attributed Document

kentucky (,), officially the commonwealth of kentucky, is a state located in the east south-central region of the united states.

Explanation: the original interlocutor relayed some information from the document and also embellished with their own personal experiences (that are not verifiable) to make a more engaging conversation.

Example 3. (QReCC)

Query

what musical has the song you'll never walk alone

Response *(Reference)*

you'll never walk alone is a show tune from the 1945 rodgers and hammerstein musical carousel.

Attributed Document

liverpool is one of the best supported clubs in the world.
the club states that its worldwide fan base includes more than 200 officially recognised club of the lfc official supporters clubs in at least 50 countries. notable groups include spirit of shankly.
the club takes advantage of this support through its worldwide summer tours, which has included playing in front of 101,000 in michigan, u.s., and 95,000 in melbourne, australia.
liverpool fans often refer to themselves as kopites , a reference to the fans who once stood, and now sit, on the kop at anfield. in 2008 a group of fans decided to form a splinter club, a.f.c. liverpool , to play matches for fans who had been priced out of watching premier league football.
the song "you'll never walk alone", originally from the rodgers and hammerstein musical carousel and later recorded by liverpool musicians gerry and the pacemakers , is the club's anthem and has been sung by the anfield crowd since the early 1960s.
it has since gained popularity among fans of other clubs around the world.
the song's title adorns the top of the shankly gates, which were unveiled on 2 august 1982 in memory of former manager bill shankly.
the "you'll never walk alone" portion of the shankly gates is also reproduced on the club's crest.

Explanation: The year "Carousel" was made (1945) cannot be attributed to the selected passage. The original interlocutor may have seen that detail elsewhere.

5.5.6 Examples. In the Appendix, we separately list textual examples rated as uninterpretable (Table A.17), interpretable but not AIS (Table A.18), or both interpretable and AIS (Table A.19). For the table-to-text task, we present examples in a more visual figure, Figure A.5, for better legibility. Common factors in marking text as "uninterpretable" include repetitive, degenerate language and ambiguous pronouns and ellipses. Additionally, some outputs are marked as uninterpretable because they are hard to understand "on their own". Whether or not a piece of text can be understood may also rely on things like commonsense and background knowledge that could vary depending on annotators' backgrounds (see Example 3 from Table A.17). Ambiguous references can also affect both interpretability and the AIS scores. In Example 2 of Table A.18, the retrieved document did not provide enough information to completely verify the response since it never refers to Ann Veneman by her full name. This is a seemingly minor detail, but annotators were often sensitive to this type of example since they could not verify whether the document was actually referring to the same entity as the model output. Another type of non-AIS output that frequently appeared in the QReCC data were cases where a model outputted a seemingly informative statement that—instead of being grounded to the document—was actually grounded to a previous conversation turn, sometimes repeating itself verbatim. Lastly, examples verify that AIS evaluations can be disentangled from other quality aspects, such as conversational relevance. This was challenging to instruct to annotators as it is instinctual to judge quality more holistically, and they were explicitly given instructions with multiple examples illustrating what types of quality aspects to ignore. In the resulting annotations, they would mark incoherent summaries or irrelevant conversational replies as AIS if they conveyed well supported information, appropriately disregarding other aspects of quality.

5.6 Comparison to Prior Work

Despite a significant amount of work devoted to hallucination across multiple NLG problems, there is no unified approach to evaluate whether system-generated statements are supported by underlying source documents. Human evaluation studies vary from paper to paper and detailed, reproducible annotation instructions are often unavailable (Belz, Mille, and Howcroft 2020). Likewise, the use of terminology for describing and defining evaluation criteria lacks consistency and further complicates replicability (Howcroft et al. 2020).

In Table 13, we summarize a selection of relevant annotation efforts and how they compare to our framework. We note that it isn't possible to perform apples-to-apples quantitative comparisons between these due to key differences in the annotation setups (annotation span; variation in tasks and datasets; metrics for capturing interannotator agreement; different labels and labeling guidelines; the effects of different pools of annotators). Instead, we underscore that our contribution is building on these prior works to present a robust, task-agnostic, reproducible annotation framework. In prior literature, these reported human annotation studies are typically task-specific (e.g., using an annotation user interface designed for summarization data only). In contrast, our annotation framework is task-agnostic by design, and we demonstrate how it can be applied to three different tasks. We address the challenges of adapting annotations to multiple tasks by explicitly formalizing the evaluation of attribution as a replicable and extendable conceptual framework. As part of our definition of attribution, we outline a more formal background for "information conveyed by the text"—in particular through the use of explicatures (see Figure 1 for examples). Prior literature also typically reports fewer interannotator quality metrics and operational statistics. In addition to reporting

Table 13
How attribution is evaluated in language model output across AbSum (Maynez et al. 2020), DialFact (Gupta et al. 2022), FEQA (Durmus, He, and Diab 2020), FRANK (Pagnoni, Balachandran, and Tsvetkov 2021), and AIS (this paper). *Span* represents the span over which human annotators provided judgments. *IAA* lists interannotator agreement metrics: percent agreement with majority (% Maj), pairwise agreement (PA), Krippendorff's alpha (Kα), Cohen's Kappa (Cκ), Fleiss Kappa (Fκ), and agreement against experts (Exp). *Process* indicates how the annotation processes were documented: the description of training and audit procedures ($^\diamond$), and performance metrics like quality and task completion times over time (*). *Instr* and *UI* indicate the availability of the instructions for human annotators and the annotation interface: fully released (+), screenshots (†), partial (*), not released (−).

Paper	Span	Task(s)	Data	IAA	Process	Instr	UI
AbSum	Phrase	*Sum*	XSUM	Fκ	−	−	† *
DialFact	Output	*Dialogue*	WoW	Kα	Training$^\diamond$, Audits$^\diamond$	+	−
FEQA	Sentence	*Sum*	CNN/DM, XSUM	% Maj	−	+	−
FRANK	Sentence	*Sum*	CNN/DM, XSUM	Fκ, % Maj, Cκ Maj Exp	Training$^\diamond$, Audits$^\diamond$	† *	†
AIS	Output	*QA*, *Sum*, *T2T*	CNN/DM, WoW, QReCC, ToTTo	Kα, Kα Exp, PA, PA Exp, F1, F1 Exp	Training$^{\diamond\star}$, Audits$^{\diamond\star}$	+	†

multiple *interannotator* quality metrics (Tables 3, 4, 5, 7, 8), we also include other quantitative studies such as how long it takes to train annotators and task completion times at different stages of the annotation (Table 9), which is important for understanding how to replicate such evaluation efforts. Lastly, whereas prior evaluations may have included limited release of instructions and annotation user interface (UI), we release full sets of instructions and screenshots of the UI (see Appendices) along with annotated data to maximize replicability.

5.7 Ablations of Annotation Task Design

In this section, we study how different design choices in the annotation framework affected the results of the human annotations. We compare with three variations of our annotation design: (1) removing the interpretability question as a filtering step; (2) replacing binary yes/no answer options with a 5-point Likert scale; (3) annotating individual sentences in model output rather than a single rating for the entire output.

5.7.1 Setup. We ran an ablation study using 200 CNN/DM summarization data examples which include the 50 that were expert annotated. We ran experiments with different sets of (untrained) annotators in three setups investigating three key design choices in the annotation task design:

- Interpretability filtering: We compare to a task design (*A1*) that removes the interpretability filtering step. In this variant, there was no interpretability question and the directions regarding interpretability were removed from the task instructions. This left the attribution question as the only question in the task.

- Binary yes/no options: We compare to a task design (*A2*) where the binary AIS answer options were replaced by a 5-point Likert scale. We included extra instructions and examples for the annotators on the differences between "fully/mostly/somewhat/mostly not/not at all" attributable. When comparing this to our original results we map the results to binary yes/no scores using either a relaxed cut-off (≥ 4) or a strict cut-off ($== 5$).
- Annotating the full output: We compare to a task design (*A3*) where the annotations occur at the sentence level instead of annotating the full model output. Annotators were first shown the full output as context and then asked to annotate each sentence separately for both interpretability and attribution. When comparing this to our original results we map the results to binary yes/no scores by defining an AIS output as one that is interpretable and attributable for all individual sentences.

We also re-ran the original task design (*Original*) with the same 200 examples, in order to avoid outside differences that may have affected the results (e.g., running at a different time of year with a different pool of annotators).

After completing the task, all annotators were asked to fill out a survey that had both multiple-choice questions (about their confidence in the task and their perceptions of the task difficulty) and free-response questions (asking which parts of the task were easiest, hardest, and most time-consuming).

5.7.2 Ablation Results. We present the results of the ablations in terms of the time taken, the differences in quality of the resulting annotations, and the insights brought up in the annotator post-task surveys.

Time-taken per Ablation. In Table 14, we present the average time taken per task as well as the median time and standard deviation. We found that the variant in which annotators annotate each sentence separately (A3) takes the longest amount of time (about a minute more than the original task design). We also found that the time taken in the task variant with a Likert scale (A2) had more variance with more outliers that took a long time. Our original task design took the shortest amount of time on average.

Comparison of Annotation Quality. We compare the results of the new ablations to the annotations from the control (the variant using the original task design). For the 50

Table 14
Ablated times: The amount of time taken per annotation for each of the task design variant in the small-scale ablation trials.

	Time per annotation (sec)		
Trial	Mean (↓)	Median (↓)	Std Dev
A1: No Interpretability Question	210.3	199.8	124.8
A2: Replace Y/N with Likert Scale	223.3	186.3	160.2
A3: Sentence-Level Annotations	258.1	248.3	115.7
Original Task Design	203.7	176.9	136.7

Table 15
McNemar's test comparing the scores of each ablation vs the control (graded against the expert annotated data). To map to the binary scale that we use in the original task design, we try two different cut-offs for the Likert study (A2)—a relaxed cut-off where AIS is anything scored ≥ 4 and a strict cut-off where where AIS is only scores $== 5$. Although we don't observe significant differences, there is a slight improvement over the control study when using the Likert scale, but only with the strict cut-off.

		A1			A2 (relaxed)			A2 (strict)			A3	
		+	−		+	−		+	−		+	−
vs. original task design	+	21	6	+	23	4	+	24	3	+	24	3
	−	5	18	−	4	19	−	7	16	−	5	18
McNemar's test p-value		1.00			0.724			0.343			0.727	

examples scored by experts, we count the annotator consensus as correct when it agrees with the expert ratings. We compare the correct/incorrect examples for each ablated version against the annotations from the original task design and use McNemar's test to study the differences (Table 15). We don't find any significant differences between the ablations and the control. Though not significant, the Likert-variant (A2) may align more with the experts by a slight margin when using a strict threshold on the AIS scores.

We also note that the model rankings found by all of these ablations and the control were the same as each other (just by different margins), meaning that any of these task designs would have produced the same conclusions about relative model performance on this output.

Annotator Usability. We present the usability survey results in Table 16. We note that the results do not show significant differences due to the small number of annotators recruited per ablation (5–10), but instead offer qualitative insights into annotators' general impressions about the task, and what might be improved. In general, the annotators tended to be most confident and have the least perceived difficulty when provided with a Likert scale (A2). They also had more confidence in their answers when there was no interpretability question (A1), though they felt the least confident of their understanding of the instructions in that case. They tended to have a worse impression of the sentence-level annotation task design (A3), expressing the most perceived difficulty and less confidence in their annotations. Annotators generally identified similar things as being easy/hard/time-consuming in the different studies. However, when there was no interpretability question, a few annotators commented that it took them longer to figure out the attribution on examples with incoherent grammar, unknown words, or insufficient context. Additionally, some commented that the examples that were partially attributable took longer to evaluate when using the Likert scale since the annotators had to spend more time discerning the differences between ratings.

5.7.3 Ablation Takeaways. Compared to our original task design, removing the interpretability question (A1) does not affect the results by much in terms of time taken or quality. It may, however, negatively impact the annotators' confidence in their understanding of the instructions. A few annotators in this study also reported that they had to spend a longer time judging attribution when the summary had incoherent grammar

Table 16
Ablation Survey Results: We asked each group of annotators to complete a post-task survey.

	Annotator impressions: Likert scores (out of 5)		
Trial	Difficulty (↓)	Understood Instructions (↑)	Confident in Responses (↑)
A1	2.1	3.9	**4.5**
A2	**1.9**	4.4	**4.5**
A3	2.8	4.4	4.2
Original	2.4	**4.5**	4.2
	Common free-response impressions		
A1	*Easiest:* reading the summary *Hardest:* reading source documents; evaluating abstractive information *Took Longest:* incoherent summaries; looking for unattributable information; lacking context; reading time		
A2	*Easiest:* easy in general; checking for attribution and interpretability *Hardest:* reading source documents; evaluating abstractive information *Took Longest:* evaluating "somewhat attributable" information; reading time; checking for attribution		
A3	*Easiest:* easy in general; checking for attribution *Hardest:* evaluating interpretability per line *Took Longest:* reading the source document; checking for attribution		
Original	*Easiest:* looking for attributable information *Hardest:* looking for unattributable information; incoherent summaries *Took Longest:* looking for unattributable or abstractive information in the source; reading the source document		

or unknown words, which may be a consequence of not asking them to evaluate this part separately as the interpretability step. Overall, compared to our original task design, removing the interpretability question may not cause much of an impact.

Splitting the attribution assignment into a sentence-by-sentence annotation (A3) increased the amount of time per annotation by about a minute. It also caused the annotators to have more perceived difficulty in the task. The annotation quality itself did not change by much. However, this does provide more fine-grained labels and is a feasible alternative way of annotating AIS for researchers who prioritize more fine-grained detail and have different constraints on time and/or cost.

The only alteration that may have improved quality over the original task design was using a 5-way Likert scale (A2). It was also the best received by annotators in terms of perceived difficulty and confidence in responses. However, there are a few caveats to this approach. It took a bit longer to complete than the original task with more outliers that took a long time (we observed that raters spent longer time on average on examples that they rated as "somewhat" or "mostly not" attributable which may have required more nuanced reading). The quality did improve on the expert-annotated portion of data but only when we used a very strict cutoff on the Likert scale, treating any score less than 5 out of 5 as unattributable. Even then, the improvements were not consistent or significant when using McNemar's test to compare to the original task design. Our

results suggest that using a Likert scale could be useful to improve annotators' overall experience although it may take them longer to evaluate certain examples. It may also improve the overall AIS ratings by a small amount, but the threshold for what is considered "AIS" has to be carefully chosen.

An additional takeaway is that, in all of the task variations, the annotators consistently commented that reading the source document took the longest time and was often seen as the most difficult part of the task. In cases where the output is fully attributable (particularly more extractive cases), they saw this as less of a bottleneck as it was easy to just pick out the parts of the document that contained the supporting information. But for abstractive or unsupported summaries, they had to read the document fully to look for the necessary information. Because reading the full document is a necessary step in determining the attribution, this can be a huge bottleneck in the task. We recommend that future work on improving this task investigate other ways of visually representing source documents (e.g., automatic highlighting of salient terms) that might help improve reading times.

6. Discussion

Generative models have been advancing toward human-like competence in some aspects. Their real-world application in consumer-focused information products is becoming more attractive—for example, for summarizing original descriptions of events, or for deriving answers to pertinent questions about the world. Traditionally, this type of information transformation has been performed by specialized human experts (e.g., journalists, researchers), who are required to meet a variety of standards of accuracy and accountability, maintaining one or more sources for a proposition and performing fact-checking. The task could also be likened to the practice of law, where norms are examined for their subsumptive relationship to a set of circumstances, and where both close reading and a set of conventionalized tests aid this determination.

We formalize a specific sub-task of fact-checking, namely, verification against a known source, as a necessary but not sufficient step in ensuring the quality of generated text. We show that with the right training, careful instructions, and optimized user interfaces, we can delegate the judgment of attribution to underlying source(s) to crowd workers, but we also find limitations. Following the data collection we described, we found it necessary to set some standards in our instructions to annotators. This includes setting expectations for named entities, for example, whether first and last names are needed to identify an individual and to link them between evidence and statement, or if a place name without qualification may be acceptable as long as there are no other well-known places of the same name. Similarly, as statements and evidence become more complex, annotators inevitably draw inferences using personal world knowledge. This is unavoidable and is inherently noisy (Pavlick and Kwiatkowski 2019); ground truth is ambiguous, just as journalists, researchers, or judges often legitimately disagree. Possible model outputs fall on a spectrum ranging from synthesized information to the mostly unassailable extractive generations (Ladhak et al. 2022). AIS does not set policy about where model output should fall: Its users still need to decide where to draw the line.

Additionally, we observe that there are more specific sub-classes of hallucinations and error types (see Section 5.5.6). Because the focus of this work is the feasibility of AIS as a score for comparing model performances, we do not distinguish between different types of error patterns. For researchers who intend to characterize attribution of a particular model, it may be useful to diagnose specific error patterns, though we

think that a separate annotation framework would be best-suited to diagnose these, which we leave to future work.

AIS is limited to propositions that can be judged with the "according to" framework. AIS is not applicable to questions (without presuppositions) or imperatives (commands and requests). There are also scenarios where strict attribution contradicts other desirable output characteristics (e.g., chit-chat systems). We did not examine AIS on such data. How to evaluate hybrid systems that mix entertaining and informative communicative goals—capturing the attribution of the informative portion but ignoring the rest—is unclear, as is the question of whether systems with blurry boundaries between what is and is not subject to attribution should exist at all. To investigate the feasibility of the AIS annotations and properly calibrate annotator responses, we have focused on tasks where the grounding source is explicitly defined, though this could be extended to more indirect grounding tasks as well (e.g., tasks that require more background knowledge).

We have purposefully limited the availability of context in our definition. Practical human–computer interactions may actually take place in context beyond the shared time t that is used in the definition (Section 3), perhaps because the communication channel is richer than a text-based conduit of transmission, and because it may be further extended by multi-session interaction history. It is important that annotators remain aware of the notion of explicature, resolving explicit references and implicit topics available to the communicators. It is possible that the use of models that perform this task (Choi et al. 2021) can improve the performance of annotators. We are also aware that this task requires close reading, which is challenging to implement on crowdsourcing platforms where speed, efficiency, and cost are incentivized instead. Again, models may be useful in extracting explicit, elementary propositions from complex statements, making this task easier for annotators.

Another approach for reducing the workload would be to use an automatic evaluation model to help score easier system outputs and only rely on humans for evaluating the more challenging ones. N-gram overlap or NLI models may work reasonably well as starting points in automatic AIS evaluation, particularly where the output is very extractive. However, such simple modeling approaches are not as feasible for evaluating very abstractive output (Dziri et al. 2022), and may not be as suitable to cases where the attributed source is long or requires extra context. Therefore, using automatic systems in conjunction with (or instead of) human annotations will require the development of more nuanced automatic AIS detection approaches. Our annotated data could be useful in evaluating the quality of such automatic AIS evaluation systems. We will examine such approaches in future work.

7. Conclusion

In this article, we define a new evaluation framework called *Attributable to Identified Sources*, which allows us to inspect whether information in generated text can be supported by source documents. We provide formal definitions of AIS and descriptions of how it can be applied to three different NLG tasks (conversational QA, summarization, and table-to-text generation). We validate this evaluation framework quantitatively on human evaluation studies, in which annotators rated the AIS of model output as part of a two-stage annotation pipeline. The results of the human evaluation studies demonstrate that high-quality AIS ratings can be obtained empirically. The results shed light on some of the ongoing challenges in training NLG models; having solid AIS is the basis for addressing them.

Appendix A. Examples of Annotations

Table A.17
Examples marked by annotators as **uninterpretable**.

Example 1. (Wizard of Wikipedia)

Conversation History

APPRENTICE: hi, can you tell me about parenting?

WIZARD: its the process of promoting and supporting the physical, emotional, social, and intellectual development of your kid, its really important for kids

APPRENTICE: thats fair enough, do you have kids?

WIZARD: no, but i am planing to have a big family one day, i really like the thought of raising a child

Query

yeah i dont have kid's but i'm really looking forward to it soon. when do you think is a good age to have kids?

Response *(T5 with Controls)*

a good parent may be referred to as a "good parent"

Rating: Not Interpretable

Example 2. (QReCC)

Query

how did the wildhearts band members meet each other?

Response *(T5-Small Pretrained)*

ginger

Rating: Not Interpretable

Example 3. (CNN/DM)

Summary *(Reference)*

daniel andersson , helsinborg 's 42-year-old kit man , kept a clean sheet .
the emergency stopper played in season opener against kalmar .
henrik larsson 's first-choice goalkeepers were both out injured .
the former goalkeeper earned one cap for sweden back in 2001 .

Rating: Not Interpretable *(Note: This is actually a reference summary. Annotators selected this as unclear as it is difficult to understand on its own. The original journalist writing this may have assumed some amount of commonsense/background knowledge about football/soccer for the intended audience as well as other information such as what games were being played that day. The interpretability judgement could vary depending on annotators' familiarity with the topic.)*

Example 4. (CNN/DM)

Summary *(Pointer-Generator Network)*

ivan carlos , 22 , and eighteen-year-old brenda avilez , 18 , were sleeping in a trailer at the lone oak mobile home in palmetto at 2.30am wednesday .
the driver of the white chevrolet suv lost control of the vehicle and crashed through the fence at the lone oak mobile home in palmetto .
avilez was nine months pregnant at the time , the herald reports , the fetus did not survive.
crawford , 21 , was taken to a hospital with non-life-threatening injuries .

Rating: Not Interpretable *(Note: There's some degenerate/repetitive language (18, eighteen) as well as unclear references that lack context ("the white chevrolet suv").)*

Table A.18
Examples marked by annotators as **interpretable but not AIS**.

Example 1. (Wizard of Wikipedia)
Conversation History
WIZARD: my pet horse is named beauty. she's domesticated :)
APPRENTICE: that must be awesome to have a pet horse! do you ride it?
WIZARD: yeah! we're not sure what breed she is, because we rescued her, but there are all sorts of different types of breeds and sizes of horses.
Query
how many different breeds are there? what color is your horse?
Response *(T5-Base)*
i'm not sure, but there are a lot of domesticated horses, but some feral ones are wild.
Attributed Document
horses in the subspecies "caballus" are domesticated, although some domesticated populations live in the wild as feral horses.
Rating: Interpretable, not AIS

Example 2. (QReCC)
Conversation History
USER:are there any other interesting aspects about ann veneman?
SYSTEM: ann veneman has received several awards and distinctions throughout her career.
Query
what kind of awards has ann veneman won?
Response *(T5-Small)*
in 2009, ann veneman received the award of distinction from the university of california davis college of agricultural and environmental sciences.
Attributed Document
In 2009, she received the award of distinction from the university of california davis college of agricultural and environmental sciences. veneman is an honorary member of rotary international (2008), received sesame workshop's leadership award for children (2006), and a humanitarian award from the united nations association of new york (2006). ...
Rating: Interpretable, Not AIS *(Note: The information is partially represented in the document but there's some referential ambiguity as the document only refers to "she" and "veneman" but it is unclear whether it's referring to "Ann Veneman" specifically. 3 of 5 annotators marked this as not AIS.)*

Example 3. (QReCC)
Conversation History
USER: when was the verve's fiirst break up?
SYSTEM: the verve's turmoil continued into 1995's a northern soul. the album reached the uk top 20 in july, but ashcroft broke up the band three months later, just before history.
USER: was history by the verve a success?
SYSTEM: history was released 18 september 1995 as the third and final single from a northern soul, charting at #24 in the uk singles chart.
Query
what other album was released by the verve besides a northen soul?
Response *(T5-Small without Evidence)*
the verve's turmoil continued into 1995's a northern soul.
Attributed Document
commercial success and second break-up (1997–1999) in early 1997, ashcroft asked mccabe to return, stating: "i got to the point where nothing other than the verve would do for me". mccabe obliged and with the new line-up in place (tong remained on guitar alongside mccabe), the group went through a "spiritual" recording process to finish their third album urban hymns which was completed by early summer. "bitter sweet symphony" the first track on their album urban hymns , "bitter sweet symphony" is renowned for its signature swirling orchestral melody. problems playing this file? see media help ...
Rating: Interpretable, Not AIS *(Note: On a surface-level, the information seems plausibly informative but is not supported in the provided document. This is actually only superficially repeating a previous conversation turn. 5 of 5 annotators marked this as not AIS.)*

Table A.19
Examples marked by annotators as **interpretable and AIS**.

Example 1. (Wizard of Wikipedia)
Conversation History
WIZARD: what do you know about the chevrolet corvette?
Response *(T5-Base)*
well, it's a sports car manufactured by chevrolet.
Attributed Document
the chevrolet corvette, known colloquially as the vette or chevy corvette, is a sports car manufactured by chevrolet.
Rating: Interpretable, AIS

Example 2. (QReCC)
Conversation History
USER: what is the british invasion in music?
SYSTEM: the british invasion was a cultural phenomenon of the mid-1960s, when rock and pop music acts from the united kingdom and other aspects of british culture, became popular in the united states and significant to the rising counterculture on both sides of the atlantic.
Query
what era did the british invasion in music happen?
Response *(T5-Base)*
pop and rock groups such as the beatles, the rolling stones, the kinks, the dave clark five, herman's hermits, the zombies, and the animals were at the forefront of the british invasion.
Attributed Document
. . . the british invasion was a cultural phenomenon of the mid-1960s, when rock and pop music acts from the united kingdom and other aspects of british culture became popular in the united states and significant to the rising "counterculture" on both sides of the atlantic. pop and rock groups such as the beatles , the rolling stones , the kinks , the dave clark five , herman's hermits , the zombies , and the animals were at the forefront of the "invasion". . . .
Rating: Interpretable, AIS *(Note: AIS is a focused metric on one dimension of quality. This response is AIS even though it fails in other dialogue qualities, i.e. relevance to previous query.)*

Example 3. (CNN/DM)
Summary
Gary Johnson took over as Cheltenham Town manager on March 30 .
He asked his players to write 'I promise to do all I can to keep Cheltenham Town in the league' on an A3 sheet of paper .
'Some signed it and meant it, and some signed it and didn't mean it,' Johnson said .
Attributed Document
Cheltenham Town have two games to preserve their Football League status - and manager Gary Johnson has revealed one of the techniques he is using to try and bring the best out of his players.
When Johnson took over as manager of the League Two club on March 30, he wrote 'I promise to do all I can to keep Cheltenham Town in the league' on an A3 sheet of paper and asked his players to put their signature on it.
'They all signed it,' Johnson said to the BBC. 'Some signed it and meant it, and some signed it and didn't mean it.
Cheltenham Town manager Gary Johnson got every payer to pledge to give his all when he took over .
Cheltenham were beaten by Northampton in their last game and have two games left to try and stay up .
'When you come to this stage of the season you need everyone to give everything for the cause,' Johnson added.
'You also need team-mates you can rely on. The lads that are here need to know they can rely on the others - and if they can't rely on some then you have to move them on.'
Cheltenham occupy 23rd in League Two and trail 22nd placed Hartlepool United and the safety places by a point.
Their final two games are against second placed Shrewsbury and 13th placed Wimbledon.
Rating: Interpretable, AIS

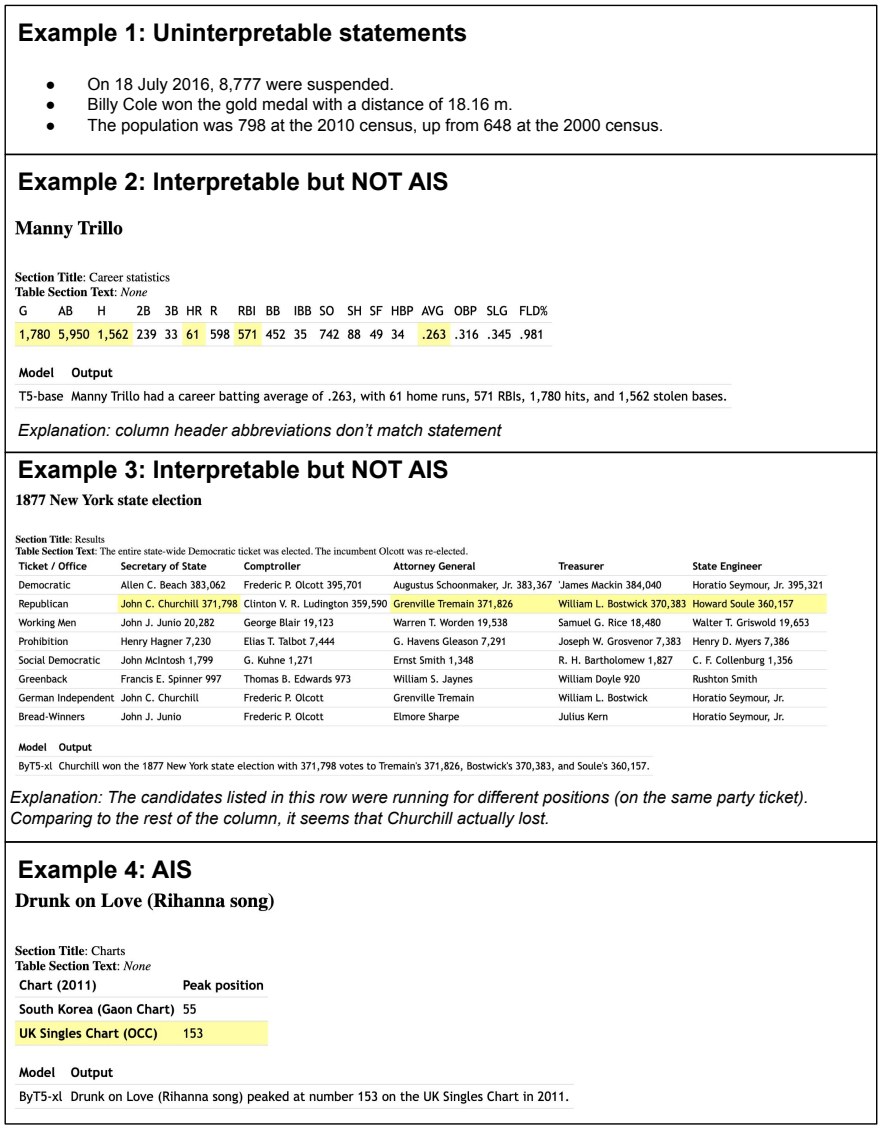


Figure A.5
Examples from the table-to-text annotations.

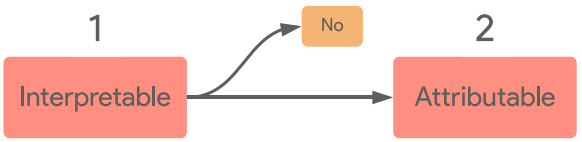
Appendix B. Evaluation Instructions for Conversational Question Answering

The following is a verbatim representation of the instructions that were presented to paid crowd annotators for performing the task alongside the interface. The prompts in the rating interface include wording from the instructions; the rating interface also contains hyperlinks to example sections in the instructions for each question and rating.

Overview

In this task you will evaluate the quality of a system-generated **response** to a **user query**. The system is trying to help the user learn about a particular topic by answering their questions. We want to rate the system response quality based on how well it represents the **original source**.

We will be using two categories to evaluate the quality of the summary: **Interpretability** and **Attribution**. You will evaluate these categories in succession. Some ratings will result in other categories being skipped. The task interface will guide you through the flow; you can also see the overall task flow in the diagram below.



Note: The system-generated responses may appear very fluent and well-formed, but contain slight inaccuracies that are not easy to discern at first glance. Pay close attention to the text. Read it carefully as you would when proofreading.

The sections below describe each of the dimensions in detail. You can also **flag** ineligible tasks; the flagging criteria are described in this section.

1. Interpretability

In this step you will evaluate whether the system response is **interpretable** by you.

You will be shown an excerpt of a conversation between a user and an assistant-like computer system. The last turn in the conversation will be a **user query** from the user followed by a **system response** attempting to respond to their query. Given the context of the user query, carefully read the system response and answer the following question:

*(Q1) Is all of the information relayed by the system response **interpretable** to you?*

This is asking whether you can understand the response. If there is some part of the response that is unclear or hard to interpret, select "No". If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. An **uninterpretable** response has diminished intelligibility due to:

- Vague or ambiguous meaning, e.g., unclear pronouns usage.
- Malformed phrases and sentences that are difficult to understand.

If the system response is interpretable by you, you will proceed to the next category. Examples of interpretability ratings and justifications are in this section.

2. Attribution

In this step, you will evaluate how well a system-generated response is **attributable** to the source document. Note that the source document is a new element that will appear in the task only when you reach this question.

Note: We refer to "**attributable to source document**" interchangeably as "**attribution**" and "**supported by the source document**". By which we mean, **all** of the information in the system response can be verified from the source document.

You will be shown an excerpt of a conversation between a user and an assistant-like computer system. The last turn in the conversation will be a **user query** from the user followed by a **system response** attempting to reply to their query. You will also be shown a document that was cited by the system as its source in attempting to answer the question (**source document**). You will use all three (user query, system response, source document) to answer the following question:

(Q2) *Is all of the information provided by the system response **fully supported** by the source document?*

This is asking whether all of the information in the system response can be attributed to the information present in the source document. If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. Attribution of a system-generated response in relation to the source document can be established by considering the following:

A. What is the **information provided by the system response?**

B. Is this information an **accurate representation of information in the source document?**

A. Definition of "the Information Provided by the System Response". Two points are key in determining the information provided by the system response:

1. The context of the system response—that is, the query and previous conversation turns—is often critical in determining the "information provided by the system response".
2. The source document should be completely ignored when determining "the information provided by the system response." (i.e., it should not be used as additional context).

Consider the following example:

User query
who plays doug and julie in days of our lives

System response
In the American daytime drama Days of Our Lives, Doug Williams and Julie Williams are portrayed by Bill Hayes and Susan Seaforth Hayes

In the above example, the meaning of the system response is clear even without seeing the query. But consider another example:

User query
who plays doug and julie in days of our lives

System response
he is played by Bill Hayes

In this case the pronoun "he" depends on the context (i.e., the query): but it is clear that the intended meaning of the system response can be paraphrased as something along the lines of "Doug Williams in days of our lives is played by Bill Hayes". In this case this paraphrase is the "information provided by the system response".

Pronouns such as he/she/it/they etc. are one case where context is needed to figure out the intended meaning of the system response. Other examples are the following (given with paraphrases of the information that is provided by the system response):

User query 1
which network is days of our lives on

System response 1
the show is on NBC

Paraphrase of information provided by system response
days of our lives is on NBC

User query 2
which network is days of our lives on

System response 2
NBC

Paraphrase of information provided by system response
days of our lives is on NBC

User query 3
how many seasons are there of days of our lives

System response 3
there are 56 seasons

Paraphrase of information provided by system response
there are 56 seasons of days of our lives

In system response 1, the phrase "the show" needs context for interpretation, but it is clear from the context of the query that it refers to "days of our lives". In system response 2, the system gives a direct answer to the query, simply "NBC", but it is clear given the query that the information provided by the system is "days of our lives is on NBC". In query 3, the phrase "56 seasons" needs context for interpretation, but given the query it is clear that the response is referring to "56 seasons of days of our lives".

In general, use your best judgment to determine the information provided by the system response. If you are unsure what the intended meaning is of the system response, make sure that you marked the example as "**No, the response is unclear.**" as part of the Interpretability stage. As one example, take the following:

User query
how many NBA championships did Michael Jordan win?

System response
it is the best team in the NBA

Paraphrase of information provided by system response
meaning is unclear

In this case it is not clear what "it" is referring to, and the meaning should be marked as being unclear. Again, use your best judgment in determining whether or not the meaning of the system response is clear.

B. Definition of "An Accurate representation of Information in the Source Document". Again, you should use your best judgment in determining whether all of the information provided by the system response is "an accurate representation of information in the source document". We give the following guidance:

- In determining this question, ask yourself whether it is accurate to say "the document says…" or "according to the document…" with the system response following this phrase. For example, is it accurate to say "according to the document below, In the American daytime drama Days of Our Lives, Doug Williams and Julie Williams are portrayed by Bill Hayes and Susan Seaforth Hayes" in the example given above?
- Be sure to check **all** of the information in the response. If only some of the information is supported in the document, but other parts of the information are missing from the document or not an accurate representation, then please mark "**No, not fully attributable**."
- The concept of "accurate representation" should be close to a journalist's conception of this phrase. For example take this excerpt from this page on Accuracy in the NPR Ethics Handbook: "When quoting or paraphrasing anyone…consider whether the source would agree with the interpretation…" In other words, if you had written the source document, consider whether you would view the system response as an accurate representation of information in that source document.

Some Final Important Notes.

Source quality. When making your judgments in this template, **do not take into account whether the underlying source document is correct or trustworthy**. This is clearly important, but will be evaluated in a separate task. The "attribution" category is used only to judge whether the information provided by the system response is an accurate representation of the underlying source document.

Context and relevance. Additionally, when rating attribution, **do not take into account the degree of relevance of the system response to the user query**. Partially and fully relevant responses can be equally assessed for attribution, regardless of how much information contained in them is relevant to the query. In both cases you should judge whether or not the system responses are an accurate representation of information in the source document even if it doesn't perfectly address the question.

Examples of attribution ratings and justifications are in this section.

Scoring and Examples

Interpretability

No, the response is unclear. The response is unclear and/or difficult to understand.

User query
when was the last time it snowed in england on christmas day

System response
it was in 2009, but it was white

Justification
The answer contains the pronoun "it" with an unclear meaning; it also sets up a contrasting statement with "but", which is hard to interpret.

User query
who sings in walk hard the dewey cox story

System response
he and john c. Reilly sang all the songs and played guitar on most of them.

Justification
The answer contains the pronoun "he" with an unclear meaning.

User query
what event marks the beginning of the great depression

System response
after a major fall in stock prices that began around september 4, 1929, and became worldwide news with the stock market crash of October 29, 1929

Justification
The answer is malformed and difficult to interpret: the sentence appears to be missing the subject somewhere near "and became".

Attribution

Yes, fully attributable. All of the information in the system response is supported by the document.

User query
when did the subway open in new york

System response
the subway was first opened in 1904

Source document
The New York City Subway is a rapid transit system owned by the City of New York and leased to the New York City Transit Authority, a subsidiary agency of the state-run Metropolitan Transportation Authority (MTA). Opened in 1904, the New York City Subway is one of the world's oldest public transit systems, one of the world's most used metro systems, and the metro system with the most stations. It offers service 24 hours per day on every day of the year.

Justification
The information in the response is accurately representing the document.

No, not fully attributable. The response contains any amount of information that is not supported by the document (including responses that are only partially or not at all supported).

User query
when did the subway open in new york

System response
the subway was first opened in late 1904.

Source document
The New York City Subway is a rapid transit system owned by the City of New York and leased to the New York City Transit Authority, a subsidiary agency of the state-run Metropolitan Transportation Authority (MTA). Opened in 1904, the New York City Subway is one of the world's oldest public transit systems, one of the world's most used metro systems, and the metro system with the most stations. It offers service 24 hours per day on every day of the year.

Justification
The document confirms that it was opened in 1904 but it doesn't specify if it was "late 1904".

User query
who has the most all star mvp awards

System response
bob pettit, kobe bryant, oscar robertson, michael jordan, shaquille o'neal, and lebron james have all won the award four times.

Source document
Bob Pettit and Kobe Bryant are the only two players to win the All-Star Game MVP four times. Oscar Robertson, Michael Jordan, Shaquille O'Neal, and LeBron James have each won the award three times.

Justification
*The cited document only supports Bob Petit and Kobe Bryant. The others listed only won the award **three** times according to the document.*

User query
who expanded the palace of versailles to its present size

System response
louis xiv, his successor, was a very good person.

Source document
Louis XIII's successor, Louis XIV, had a great interest in Versailles. He settled on the royal hunting lodge at Versailles, and over the following decades had it expanded into one of the largest palaces in the world. Beginning in 1661, the architect Louis Le Vau, landscape architect André Le Nôtre, and painter-decorator Charles Lebrun began a detailed renovation and expansion of the chteau. This was done to fulfill Louis XIV's desire to establish a new centre for the royal court. Following the Treaties of Nijmegen in 1678, he began to gradually move the court to Versailles. The court was officially established there on 6 May 1682.

Justification
The answer is correct but the last part (he was a "very good person") is an opinion and not supported by the document.

Flag

There is a flag button in the bottom left corner of the task interface. Once flagged, you can proceed onto the next task. Use it to report tasks that are ineligible for reasons such as:

1. Missing components in the task. Some tasks may have missing user queries, responses, or source text. They should be flagged.

2. Malformed text. Some text may be severely malformed with unintelligible artifacts (e.g. html code, unformatted tables, etc.). If any component of the task contains malformed

text, the task should be flagged. The example below shows various types of malformed text.

User query
where did remember the titans camp take place

System response
i think it took place in a gymnasium in san diego, where the black students auditioned to play for the $team_winner until boone arrives, but the meeting turns into a fiasco when yoast and white students interrupt.

Source document
The black students have a meeting in the gymnasium in auditioning to play for the team until &&&Boone&&&& arrives, but the meeting turns into a fiasco when Yoast and white students interrupt. On August 15, 1971, the players gather and journey to Gettysburg College, where their training camp takes place.

3. Source document is underspecified. Some source documents may not have sufficient information to determine whether it does/doesn't support the response. They may be too short and/or lack critical information that would be necessary to rate the response.

User query
where is the hero golf challenge being played

System response
the event will be held at the Albany development in the Bahamas

Source document
Beginning in 2015, the event will be played at the Albany development in the Bahamas, for at least three years.

Justification
The source document does not specify what event it is referring to, and so it's impossible to determine whether this is referring to the same "hero golf challenge" or not . Without the explicit mention of "hero gold challenge", the task should be flagged.

User query
who is known as the super fast boy in the series the icredible

System response
spencer fox as dashiell robert parr

Source document
Spencer Fox as Dashiell Robert Parr / Dash, the Parrs' second child, who possesses super-speed

Justification
It's unclear whether the source document is referring to the same character as the one in the query because it does not specify the series ("the incredible") or the name of the character ("the super fast boy").

4. Source document is difficult to understand because it requires expert-level knowledge. Some documents may include scientific formulas, obscure terminology, etc. If you can still understand enough of the document to rate the attribution, please do so. But, on the other hand, if properly evaluating the response requires expertise in a particular area, please flag it.

User query
what is the rate limiting enzyme of kreb's cycle

System response
kreb is the enzyme that uses fad as a prosthetic group

Source document
Substrates Products Enzyme Reaction type Comment 0 / 10 Oxaloacetate + Acetyl CoA + H2O Citrate + CoA-SH Citrate synthase Aldol condensation irreversible, extends the 4C oxaloacetate to a 6C molecule 1 Citrate cis-Aconitate + H2O Aconitase . . .

Justification
In order to be able to evaluate whether this source document supports the response, it requires a deeper understanding of scientific equations and terminology contained in the document. Because this example requires scientific expertise to evaluate it properly, it should be flagged.

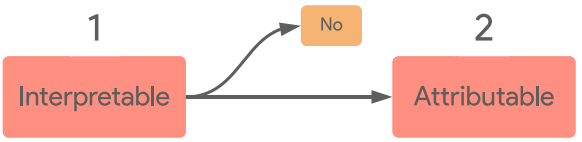
Appendix C. Evaluation Instructions for Summarization

The following is a verbatim representation of the instructions that were presented to paid crowd annotators for performing the task alongside the interface. The prompts in the rating interface include wording from the instructions; the rating interface also contains hyperlinks to example sections in the instructions for each question and rating. The summarization instructions were developed after the conversational QA instructions had been established and the annotators had been trained on the conversation QA task.

Overview

In this task you will evaluate the quality of a system-generated **summary**. The system's goal is to summarize the **source news article**, while remaining truthful to it. We want to rate the quality of the summary based on how well it represents the original source.

We will be using two categories to evaluate the quality of the summary: **Interpretability** and **Attribution**. You will evaluate these categories in succession. Some ratings will result in other categories being skipped. The task interface will guide you through the flow; you can also see the overall task flow in the diagram below.



Note: The system-generated summaries may appear very fluent and well-formed, but contain slight inaccuracies that are not easy to discern at first glance. Pay close attention to the text. Read it as carefully as you would when proofreading.

The sections below describe each of the dimensions in detail. You can also **flag** ineligible tasks; the flagging criteria are described in this section.

1. Interpretability

In this step you will evaluate whether the system summary is **interpretable** by you.

You will be shown a system-generated **summary** of a news article. Note that the news article from which the summary is derived is hidden at this step, because we need

to evaluate whether the summary is *interpretable on its own*. Carefully read the summary and answer the following question:

*(Q1) Is all of the information relayed by the system summary **interpretable** to you?*

This is asking whether you can understand the summary on its own. If there is any part of the summary that is unclear or hard to interpret, select "No". If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. An **uninterpretable** summary has diminished intelligibility due to:

- Vague or ambiguous meaning, e.g., unclear noun references or pronouns usage.
- Malformed phrases and sentences that are difficult to understand.

If the summary is interpretable by you, you will proceed to the next category.
In the section below, we show in more detail the kind of reasoning that should be used for establishing interpretability of summaries. More examples of interpretability ratings and justifications are in this section of the appendix.

Interpreting the information provided in the system summary. Consider the following example:

> **Summary**
> seismologists put the magnitude at 7.9 for an earthquake that hit kathmandu today, which would actually make it about 40% larger than the 7.8 currently being reported .

In the above example, the meaning of the summary is clear even without seeing the original news article. It is clear what the summary is reporting on and it stands on its own, that is, this summary is **interpretable**. It should be marked as "**Yes, I understand it.**" But consider another example:

> **Summary**
> seismologists put the magnitude at 7.9 , which would actually make it about 40% larger than the 7.8 currently being reported .

In this case the meaning of the phrase "the magnitude" obviously depends on some context, but that context is missing in the summary. Without additional information that clarifies that the magnitude refers to an earthquake that occurred in a specific location (Kathmandu), the summary is **difficult to interpret** and it does not stand on its own. It should be marked as "**No, the summary is unclear.**"
Noun phrases that require clarifications of this kind are one case where interpretability can be diminished. Other examples include nouns and pronouns without a (clear) reference and malformed phrases and sentences:

> **Summary 1**
> the project is hoped to open at the former arts centre , la llotja . friend and producer jaume roures of mediapro is leading the tribute . the museum follows allen 's vicky cristina barcelona, set in the city

Summary 2
new england patriots tight end aaron hernandez has pleaded not guilty to murder and two weapons charges . he 's accused of orchestrating the shooting death of odin lloyd . it 's scheduled to begin in may , but not legally required to get a conviction .

Summary 3
john stamos announced monday night on " jimmy kimmel live " . the show will feature candace cameron bure , who played eldest daughter d.j . tanner in the original series , which aired from 1987 to 1995 , will both return for the new series .

In summary 1, the phrase "the project" needs context for interpretation ("*what project is being reported on?*"). Likewise, "the museum" and "the city" are unclear ("*what museum is this?*", "*what city is this taking place in?*") In summary 2, the system provides a clear reference for the pronoun "he" ("hernandez"), but a reference for the pronoun "it" is missing ("*what is scheduled to begin in may?*"). Also it is not clear what "not legally required to get a conviction" is referring to. In summary 3, the first sentence is malformed because it is missing the announcement ("*what did john stamos announce?*"). The second sentence is difficult to understand ("*who does 'both' refer to?*", "*who will return to the new series?*").

In general, use your best judgment to determine the information provided by the summary. If you are unsure what the intended meaning of the summary is, err on the side of marking it with "**No, the summary is unclear.**"

2. Attribution

In this step, you will evaluate how well a system-generated summary is **attributable** to the source news article. Note that the source news article is a new element that will appear in the task only when you reach this question.

Note: We refer to "**attributable to the source news article**" interchangeably as "**attribution**" and "**supported by the source news article**". By which we mean, **all** of the information in the system-generated summary can be verified from the source news article.

You will be shown a system-generated **summary** of a news article. You will also be shown the news article that was used by the system to generate this summary (**source news article**). You will use both of these to answer the following question:

(Q2) *Is all of the information provided by the system summary **fully supported** by the source document?*

This is asking whether all of the information in the system summary can be attributed to the information present in the source news article. If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. A **fully supported** (or **attributable**) system-generated summary contains an accurate representation of information in the source news article. No information in the summary is unattested when compared against the source news article.

In the section below, we show in more detail the kind of reasoning that should be used for establishing attribution of summaries. More examples of attribution ratings and justifications are in this section of the appendix.

Assessing the accuracy of the information in the summary against the original news article. Again, you should use your best judgment in determining whether all of the information provided by the system summary is "an accurate representation of information in the source news article". We give the following guidance:

- In determining this question, ask yourself whether it is accurate to say "the provided news article says…" or "according to the news article…" with the system summary following this phrase.
- Be sure to check **all** of the information in the summary. If only some of the information is supported in the news article, but other parts of the information are missing from the news article or not an accurate representation, then please mark "**No, not fully attributable**."
- The concept of "accurate representation" should be close to a journalist's conception of this phrase. For example take this excerpt from this page on Accuracy in the NPR Ethics Handbook: "When quoting or paraphrasing anyone…consider whether the source would agree with the interpretation…" In other words, if you had written the source document, consider whether you would view the summary as an accurate representation of information in that source document.

Some Final Important Notes.

Source quality. When making your judgments in this template, **do not take into account whether the underlying source news article is correct or trustworthy**. This is clearly important, but will be evaluated in a separate task. The "attribution" category is used only to judge whether the information provided by the system summary is an accurate representation of the underlying source news article.

Examples of attribution ratings and justifications are in this section.

Scoring and Examples

Interpretability

No, the summary is unclear. The summary is unclear and/or difficult to understand.

Summary
The bill would prevent adolescents from smoking, buying or possessing both traditional and electronic cigarettes . The bill includes a $10 fine for first-time offenders. Subsequent violations would lead to a $50 fine or mandatory community service . Dozens of local governments have similar bans, including Hawaii County and New York City .

Justification
The summary revolves around the noun phrase "the bill" that doesn't have a clear reference (what bill is it referring to?); it makes the summary difficult to understand fully.

Summary
The former England rugby star tweeted "Holy s***, I'm lost for words and emotions. All I can say is yes the dude!!!!!" After he bought the horse Zara called him an idiot, but she was there with him, cheering the horse home . Monbeg Dude was given the all-clear to run in the Grand National at Aintree after a poor showing at the Cheltenham Festival .

Justification
The answer contains the noun phrase "the former England rugby star" that lacks a clear reference (who is it?). Furthermore, the usage of pronouns "he", "him" and "she" is too vague (who are they referring to in the summary?).

Summary
john stamos announced monday night on " jimmy kimmel live " . the show will feature candace cameron bure , who played eldest daughter d.j . tanner in the original series , which aired from 1987 to 1995 , will both return for the new series .

Justification
The answer is malformed and difficult to interpret. The first sentence is missing the object of "announced" (what did john stamos announce?). Additionally, "will both return for the new series" appears to be poorly connected to the previous context.

Attribution

Yes, fully attributable. All of the information in the system summary is supported by the document.

Summary
Convicted murderer Nikko Jenkins, 28, tried to carve '666' into his forehead . But in a phenomenal case of idiocy, he used a mirror - so the numbers came out backwards . The symbol is described in the biblical book of Revelation as 'the sign of the beast', and has since been popularized by the horror movie The Omen . Jenkins was jailed exactly one year ago for shooting dead four people in 10 days after being released from prison .

Original news article
It was meant to be the ultimate symbol of menace: carving '666' into his forehead.

But in a phenomenal case of idiocy, convicted murderer Nikko Jenkins used a mirror - so the numbers came out backwards.

The symbol is described in the biblical book of Revelation as 'the sign of the beast', and has since been popularized by the horror movie The Omen.

However, with a series of upside-down 9s, Jenkins has fashioned himself an entirely unique - and irreversible - engraving.

Botched: Nikko Jenkins (pictured in 2014) recently tried to carve '666' into his forehead but did it backwards .

According to Omaha.com, Jenkins told his attorney about the incident in a phone call from his cell in Omaha, Nebraska.

It comes amid the 28-year-old's ongoing appeal that he is mentally unstable and therefore ineligible to face the death penalty.

Jenkins was jailed exactly one year ago for shooting dead four people in 10 days after being released from prison.

During his murder trial in Douglas County, Jenkins was assessed by a doctor who concluded that he was 'a psychopath' and 'one of the most dangerous people' he had ever encountered.

Psychopath': The 28-year-old, who a doctor described as 'one of the most dangerous people' he had ever encountered, may use the botched case of self-mutilation as evidence he is mentally unstable .

Jenkins pleaded not guilty, then guilty, then ineligible for trial on the grounds of insanity. However, a judge dismissed the appeals and he was sentenced to life.

The decision of whether he would be sentenced to death was delayed after Jenkins revealed he had carved a swastika into his skin.

Following months of delays, he will face a panel in July to decide his fate.

It is believed Jenkins may use his latest botched case of self-mutilation as further evidence that he is mentally unstable.

Justification
The information in the summary accurately represents the information in the source news article.

No, not fully attributable:. The summary contains any amount of information that is not supported by the source new article (including summaries that are only partially or not at all supported).

Summary
saracens lost 13-9 to clermont at stade geoffroy-guichard on saturday . the sarries pack contained five english-qualified forwards . saracens ' millionaire chairman nigel wray wants the salary cap scrapped .

Original news article
saracens director of rugby mark mccall lauded his young guns after their latest european heartache before declaring he has no intention of overspending in a competitive post-world cup transfer market .

mccall watched his side , which contained five english-qualified forwards in the starting pack , battle in vain before losing 13-9 to the clermont on saturday .

saracens ' millionaire chairman nigel wray spent much of last week repeating his belief the cap should be scrapped in order for saracens to compete at europe 's top table , raising expectations they could be set to land a ' marquee ' player from outside the league whose wages would sit outside next season 's # 5.5 m cap.

However, with a series of upside-down 9s, Jenkins has fashioned himself an entirely unique - and irreversible - engraving.

maro itoje (second left) was one of five england-qualified forwards in the saracens pack that faced clermont mako vunipola tries to fend off clermont lock jamie cudmore during a ferocious contest saracens director of rugby mark mccall saw his side come agonisingly close to reaching the final but mccall said : ' we know where we 'd like to improve our side and we 're prepared to wait for the right person . we do n't want to jump in and get " a name " just because he 's available post-world cup .

' the fact our pack is as young as it is is incredibly exciting for us . they could be the mainstay of the club for the next four to five seasons . '

billy vunipola (left) , jim hamilton and itoje leave the field following their 13-9 loss against clermont

Justification
The summary includes unattributed information: "at stade geoffroy-guichard", and the reference to the opposing team as the "the sarries", which is not supported in the original new article.

Flag

There is a flag button in the bottom left corner of the task interface. Once flagged, you can proceed onto the next task. Use it to report tasks that are ineligible for reasons such as:

1. Missing components in the task. Some tasks may have missing summaries or news articles. They should be flagged.

2. *Malformed text.* Some text may be severely malformed with unintelligible artifacts (e.g. html code, unformatted tables, etc.). If any component of the task contains malformed text that makes it difficult for you to accomplish the task, the task should be flagged.

3. *Source document is difficult to understand because it requires expert-level knowledge.* Some documents may include scientific formulas, obscure terminology, etc. If you can still understand enough of the document to rate the attributability, please do so. But, on the other hand, if properly evaluating the summary requires expertise in a particular area, please flag it.

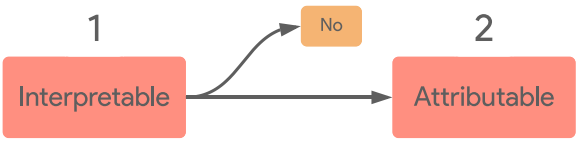
Appendix D. Evaluation Instructions for Table-to-Text

The following is a verbatim representation of the instructions that were presented to paid crowd annotators for performing the task alongside the interface. The prompts in the rating interface include wording from the instructions; the rating interface also contains hyperlinks to example sections in the instructions for each question and rating. The table-to-text instructions were developed after the conversational QA and summarization instructions had been established and the annotators had been trained on the conversation QA and summarization tasks.

Overview

In this task you will evaluate the quality of a system-generated **caption** for highlighted parts of a table. The system is trying to convert the information in the table into a natural language description (what we are referring to as the "system-generated caption"). We want to rate the quality of the system-generated caption based on how well it represents information from the **source table**.

We will be using two categories to evaluate the quality of the caption: **Interpretability** and **Attribution**. You will evaluate these categories in succession. Some ratings will result in other categories being skipped. The task interface will guide you through the flow; you can also see the overall task flow in the diagram below.



Note: The system-generated captions may appear very fluent and well-formed, but contain slight inaccuracies that are not easy to discern at first glance. Pay close attention to the text. Read it carefully as you would when proofreading.

The sections below describe each of the dimensions in detail. You can also **flag** ineligible tasks; the flagging criteria are described in this section.

1. Interpretability

In this step you will evaluate whether the system caption is **interpretable** by you.

You will be shown a system-generated **caption** of a table. Note that the table from which the caption is derived is hidden at this step, because we need to evaluate

whether the caption is *interpretable on its own*. Carefully read the caption and answer the following question:

*(Q1) Is all of the information relayed by the system caption **interpretable** to you?*

This is asking whether you can understand the caption. If there is some part of the caption that is unclear or hard to interpret, select "No". If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. An **uninterpretable** caption has diminished intelligibility due to:

- Vague or ambiguous meaning, e.g., unclear noun references or insufficient context.
- Malformed phrases and sentences that are difficult to understand.

If the system caption is interpretable by you, you will proceed to the next category.

In the section below, we show in more detail the kind of reasoning that should be used for establishing interpretability of captions. More examples of interpretability ratings and justifications are in this section of the appendix.

Interpreting the information provided in the system caption. Consider the following example:

Caption
Mikhnevich and Avdeyeva both finished with 19.66 meters in the 2009 World Championships in Athletics—Women's shot put.

In the above example, the meaning of the caption is clear even without seeing the source table. It is clear what the caption is reporting on and it stands on its own; that is, this caption is **interpretable**. It should be marked as "**Yes, I understand it.**" But consider another example:

Caption
Mikhnevich and Avdeyeva finished with 19.66 metres.

In this case the meaning of the caption obviously depends on some context ("*what did they finish?*"), but that context is missing in the caption. Without additional information that clarifies that this is a result of a sports competition, the caption is **difficult to interpret** and it does not stand on its own. It should be marked as "**No, the caption is unclear.**"

Captions that require clarifications of this kind are one case where interpretability can be diminished. Other examples include nouns without a (clear) reference and malformed phrases and sentences:

Caption 1
There are 87,814 Albanians, 624 Serbs and 361 Roma.

Caption 2
J. Thomas Heflin (D) was a member until November 1, 1920 and William B. Bowling (D) succeeded him from December 14, 1920.

Caption 3
George A. Gillett was a New Zealand dual-code international rugby.

In caption 1, the numbers are missing the context of what they are being reported on ("*what location or event do these numbers represent?*"). In caption 2, the phrase "a member" lacks a specifying reference ("*what was Thomas Heflin a member of?*"). In caption 3, the sentence is difficult to understand as it appears to be missing a noun ("*was George A. Gilet a rugby **player**?*").

In general, use your best judgment to determine the information provided by the caption. If you are unsure what the intended meaning of the caption is, err on the side of marking it with "**No, the caption is unclear.**"

2. Attribution

In this step, you will evaluate how well a system-generated caption is **attributable** to the source table. Note that the source table is a new element that will appear in the task only when you reach this question.

Note: We refer to "**attributable to source table**" interchangeably as "**attribution**" and "**supported by the source table**". By which we mean, **all** of the information in the system caption can be verified from the source table.

You will be shown a system-generated **caption**. You will also be shown a table and its associated descriptions: title, section title, and table section text. These elements provide additional context for understanding the information in the table. Finally, some cells in the table will be highlighted as helpful hints for which parts of the table are the focus of the caption. The table, descriptions, and highlighted cells (**source table**) were used by the system to create the caption. You will use all of these elements to answer the following question:

(Q2) *Is all of the information provided by the system caption **fully supported** by the source table?*

This is asking whether all of the information in the system caption can be attributed to the information present in the source table. If prompted by the interface, enter a succinctly detailed justification of your rating.

Definition. A **fully supported** (or **attributable**) system-generated caption contains an accurate representation of information in the source table. No information in the caption is unattested when compared against the source table and its associated descriptions (title, section title, and table section text).

In the section below, we show in more detail the kind of reasoning that should be used for establishing attribution of captions. More examples of attribution ratings and justifications are in this section of the appendix.

Assessing the accuracy of the information in the caption against the source table. Again, you should use your best judgment in determining whether all of the information provided by the system caption is "an accurate representation of information in the source table". We give the following guidance:

- In determining this question, ask yourself whether it is accurate to say "the provided table says" or "according to the table" with the system caption following this phrase.

- Be sure to check **all** of the information in the caption. If only some of the information is supported in the table, but other parts of the information are missing from the table or not an accurate representation, then please mark "**No, not fully attributable.**"
- The concept of "accurate representation" should be close to a journalist's conception of this phrase. For example take this excerpt from this page on Accuracy in the NPR Ethics Handbook: "When quoting or paraphrasing anyone. . . consider whether the source would agree with the interpretation. . . " In other words, if you had written the source document, consider whether you would view the caption as an accurate representation of information in that source document.

Some Final Important Notes.

Source quality. When making your judgments in this template, **do not take into account whether the underlying source table is correct or trustworthy**. This is clearly important, but will be evaluated in a separate task. The "attribution" category is used only to judge whether the information provided by the system caption is an accurate representation of the underlying source table.

Highlighted cells. Some of the cells in the table are highlighted. The highlighted cells are intended to be the focus of the caption, and can be used as a helpful hint of where to look in the table for information in the caption—though some captions may also refer to information from elsewhere in the table. **If the caption does not capture the information in the highlighted cells, but otherwise accurately represents the information elsewhere in the table and its description, please still mark it "Yes, fully attributable."**

Examples of attribution ratings and justifications are in this section.

Scoring and Examples

Interpretability

No, the caption is unclear:. The caption is unclear and/or difficult to understand.

Caption
Bradman scored 299.

Justification
The caption lacks sufficient context ("what did Bradman score in?").

Caption
Amlogic Quad-core Cortex-A53, Mali-450 MP5 OpenGL ES 2.0, and H.264.

Justification
The caption is malformed and difficult to interpret: the sentence appears to be missing a verb.

Caption
The number one album of the year was Watch the Throne by Jay-Z and Kanye West, which sold 436,000 copies, and Lupe Fiasco's Lasers, which sold 204,000 copies.

Justification
The caption is malformed because it has two albums listed as "number one album of the year". Additionally, "the year" lacks a reference ("what year was it?").

Attribution

Yes, fully attributable:. All the information in the caption is supported by the table and its description.

Caption
The first-week sales of the album The Watch the Throne by Jay-Z and Kanye West sold 436,000 copies, while Lupe Fiasco's Lasers sold 204,000 copies in the first week.

Source table

2011 in hip hop music

Section Title: Highest first-week sales
Table Section Text: *None*

Number	Album	Artist	1st-week sales	1st-week position	Refs
1	Tha Carter IV	Lil Wayne	964,000	1	
2	Take Care	Drake	631,000	1	
3	Watch the Throne	Jay-Z & Kanye West	436,000	1	
4	Thug Motivation 103: Hustlerz Ambition	Young Jeezy	233,000	3	
5	Cole World: The Sideline Story	J. Cole	217,000	1	
6	Lasers	Lupe Fiasco	204,000	1	
7	Rolling Papers	Wiz Khalifa	197,000	2	
8	Hell: The Sequel	Bad Meets Evil	171,000	1	
9	Ambition	Wale	164,000	2	
10	Blue Slide Park	Mac Miller	144,000	1	

Justification
The information in the caption accurately represents the source table.

Caption
Albanians are the largest ethnic group in Gjilan, followed by Serbs with 624 and Roma with 361 persons.

Source table

Gjilan

Section Title: Ethnic groups
Table Section Text: The vast majority of the population is Albanian, followed by Serbs, and a small number of minorities. There are among others, 978 Turks or 1% of the municipal population.

Ethnic composition

Year/population	Albanians	%	Serbs	%	Roma	%	Others	%	Total
1953	24,797	50.87	19,196	39.32					48,748
1961	29,942	57.12	18,297	34.91	735	1.50			52,415
1971	43,754	64.45	20,237	29.81	1,824	2.69			67,893
1981	59,764	71.08	19,212	22.85	3,347	3.98	1,762	2.1	84,085
1991	79,357	76.54	19,370	18.68	3,477	3.4	1,471	1.4	103,675
1998	94,218	79.4	19,481	16.4	3,568	3	1,387	1.2	118,654
2011	87,814	97,45	624	0,7	361	0,4	1,379	1,52	90,178

Source: Yugoslav population censuses for data through 1991,and Kosovo 2011 census.

Justification
The caption accurately reflects information from the source table. The fact that the Albanian ethnic group is the largest can be easily reasoned from the information in the table.

No, not fully attributable:. The caption cannot be fully attributed to the source table and its description (including captions that are only PARTIALLY or NOT AT ALL supported).

Caption
The first-week sales of the album Watch the Throne by Jay-Z & Kanye West and Lupe Fiasco were 204,000 and the first-week sales of the album Lasers by Lupe Fiasco were 436,000.

Source table

2011 in hip hop music

Section Title: Highest first-week sales
Table Section Text: *None*

Number	Album	Artist	1st-week sales	1st-week position	Refs
1	Tha Carter IV	Lil Wayne	964,000	1	
2	Take Care	Drake	631,000	1	
3	Watch the Throne	Jay-Z & Kanye West	436,000	1	
4	Thug Motivation 103: Hustlerz Ambition	Young Jeezy	233,000	3	
5	Cole World: The Sideline Story	J. Cole	217,000	1	
6	Lasers	Lupe Fiasco	204,000	1	
7	Rolling Papers	Wiz Khalifa	197,000	2	
8	Hell: The Sequel	Bad Meets Evil	171,000	1	
9	Ambition	Wale	164,000	2	
10	Blue Slide Park	Mac Miller	144,000	1	

Justification
The artists on the album Watch the Throne are only Jay-Z and Kanye West, excluding Lupe Fiasco. The first week's sales of Watch the Throne were 436,000, not 204,000. The first week's sales of Lasers were 204,000, not 436,000.

Caption
Juan Mora Fernández was the Head of State of Costa Rica, winning 11 seats in the San José, 8 in Cartago, 8 in Heredia, 3 in Escazú, 2 in Ujarrás, 1 in Térraba and 1 in Bagaces.

Source table

1825 Costa Rican Head of State election

Section Title:
Table Section Text: *None*

Candidate	San José	Cartago	Heredia	Alajuela	Escazú	Ujarrás	Térraba	Bagaces
Juan Mora Fernández	11	8	8		3	2	1	1
Mariano Montealegre				8				

Justification
The table does not have a clear identification of the reported numbers, while the caption identifies them as "seats", which is not attributable anywhere in the table or its descriptions.

Flag

There is a flag button in the bottom left corner of the task interface. Once flagged, you can proceed onto the next task. Use it to report tasks that are ineligible for reasons such as:

1. Missing components in the task. Some tasks may have missing summaries or news articles. They should be flagged.

Note that table title, section title, or table section text could be empty or designated with "None". These are acceptable and should not be flagged. See an example of an acceptable table description below:

1825 Costa Rican Head of State election

Section Title:
Table Section Text: *None*

2. Malformed text. Some text may be severely malformed with unintelligible artifacts (e.g. html code, unformatted tables, etc.). If any component of the task contains malformed text, the task should be flagged.

Caption
The lowest temperature recorded in Porto Alegre was ?? 0.3°C (31.5°F).

3. Source table is difficult to understand because it requires expert-level knowledge. Some tables may include scientific formulas, obscure terminology, etc. If you can still understand enough of the table to rate its attributability, please do so. But if properly evaluating the response requires expertise in a particular area, please flag it.

Caption
The longest-lived isotope is 18mF with a half-life of 162 ns.

Source table

Isotopes of fluorine

Section Title: List of isotopes
Table Section Text: Bold for stable isotopes Has medicinal uses

nuclide symbol	Z(p)	N(n)	isotopic mass (u)	half-life	decay mode(s)	daughter isotope(s)	nuclear spin and parity	representative isotopic composition (mole fraction)
	excitation energy							
14F	9	5	14.03432(4)	500(60)×10−24 s [910 keV]	p	13O	2-	
15F	9	6	15.017785(15)	1.1(0.3)×10−21 s [1.0(2) MeV]	p	14O	1/2+	
16F	9	7	16.011466(9)	11(6)×10−21 s [40(20) keV]	p	15O	0−	
17F	9	8	17.00209524(27)	64.370(27) s	β+	17O	5/2+	
18F	9	9	18.0009373(5)	109.739(9) min	β+ (96.86%)	18O	1+	
					EC (3.14%)	18O		
18mF	1121.36(15) keV			162(7) ns	IT	18F	5+	
19F	9	10	18.9984031629(9)	Stable			1/2+	1.0000
20F	9	11	19.99998125(3)	11.163(8) s	β−	20Ne	2+	
21F	9	12	20.9999489(19)	4.158(20) s	β−	21Ne	5/2+	
22F	9	13	22.002999(13)	4.23(4) s	β− (89%)	22Ne	(4+)	
					β−n (11%)	21Ne		
23F	9	14	23.00353(4)	2.23(14) s	β− (86%)	23Ne	5/2+	
					β−n (14%)	22Ne		
24F	9	15	24.00810(10)	384(16) ms	β− (94.1%)	24Ne	3+	
					β−n (5.9%)	23Ne		
25F	9	16	25.01217(10)	80(9) ms	β− (76.9%)	25Ne	(5/2+)	
					β−n (23.1%)	24Ne		
26F	9	17	26.02002(12)	8.2(9) ms	β− (86.5%)	26Ne	1+	
					β−n (13.5%)	25Ne		
26mF	643.4(1) keV			2.2(1) ms	IT (82%)	26F	(4+)	
					β−n (12%)	25Ne		
					β−n (6%)	26Ne		
27F	9	18	27.02732(42)	4.9(2) ms	β−, n (77%)	26Ne	5/2+#	
					β− (23%)	27Ne		
28F	9	19	28.03622(42)	46×10−21 s	n	27F		
29F	9	20	29.04310(56)	2.5(3) ms	β−, n (60%)	28Ne	5/2+#	
					β− (40%)	29Ne		
31F	9	22	31.06027(59)#	1# ms [>260 ns]	β−	31Ne	5/2+#	

Justification
In order to be able to evaluate whether this table supports the caption, it requires a deeper understanding of scientific equations and terminology contained in the table. Because this example requires scientific expertise to evaluate it properly, it should be flagged.

Appendix E. Annotation User Interface for Conversational QA Tasks

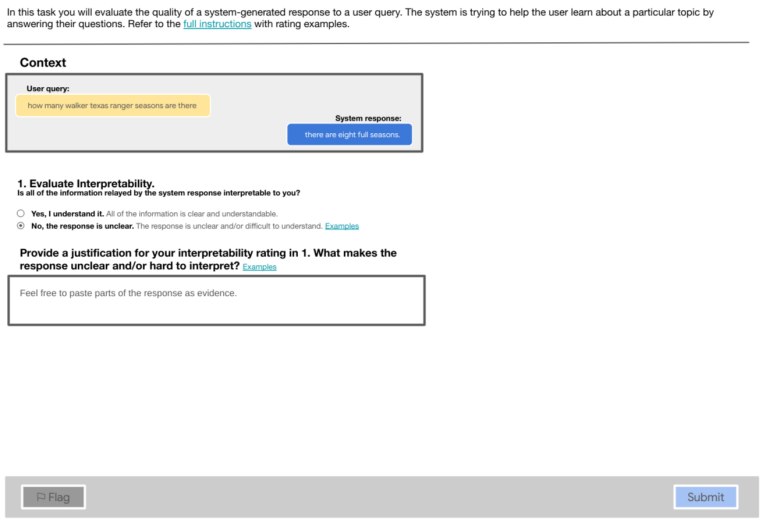


Figure E.6
Interpretability stage. The full conversation history is shown, while the source document is hidden. During training and the pilot the justification element is shown, but only if the task is rated as not interpretable.

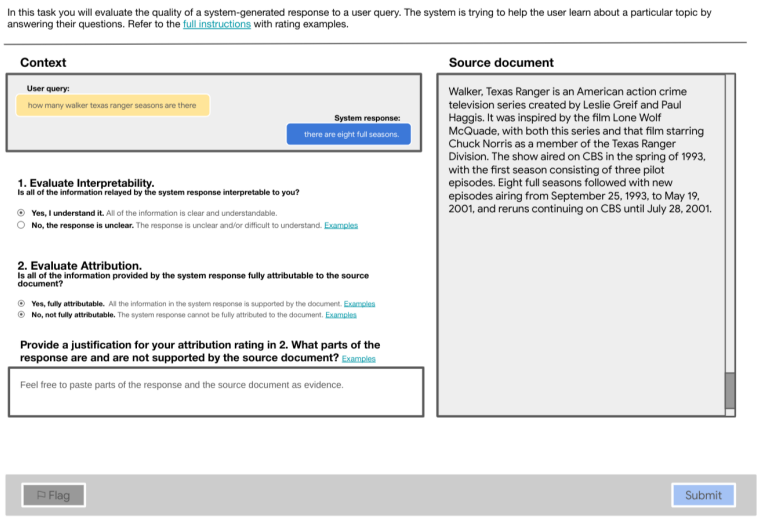


Figure E.7
Attribution stage. The source document is shown. During training and the pilot the justification element is required for all ratings. If the task is rated as not interpretable at the previous stage, the attribution stage is skipped and the annotator proceeds to the next task in the queue.

Appendix F. Annotation User Interface for Summarization Tasks

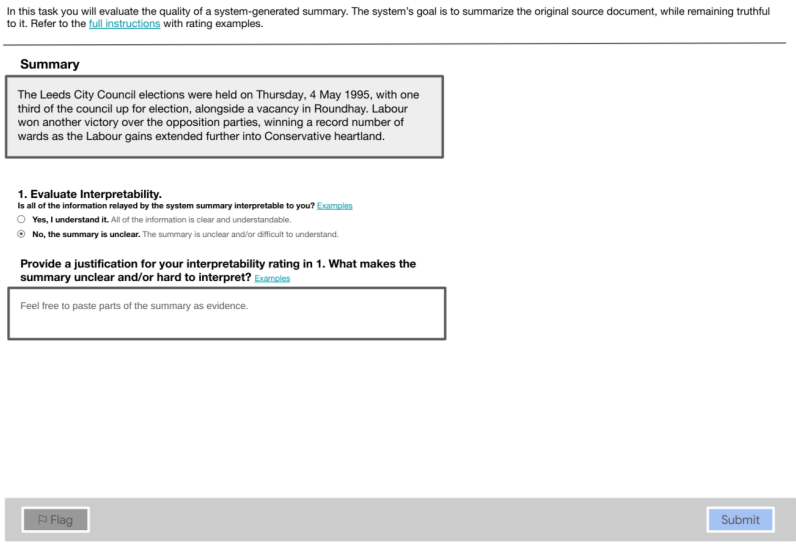


Figure F.8
Interpretability stage. The source document is hidden. During training and the pilot the justification element is shown, but only if the task is rated as not interpretable.

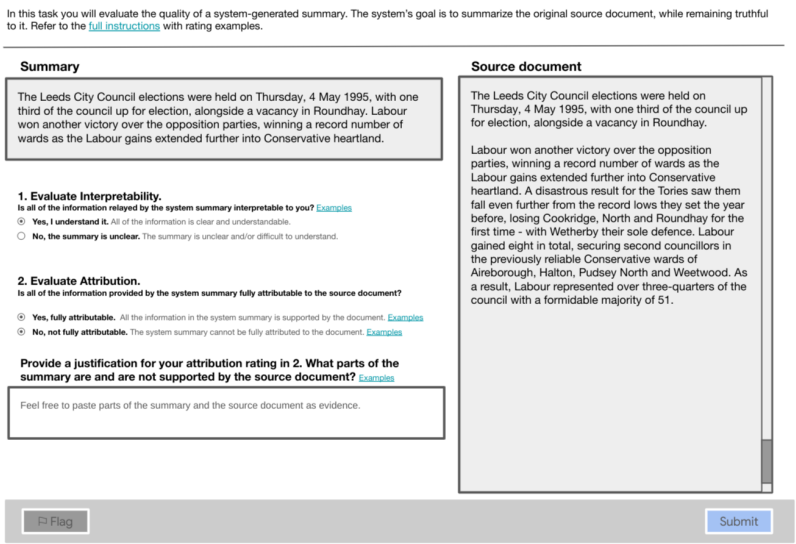


Figure F.9
Attribution stage. The source document is shown. During training and the pilot the justification element is required for all ratings. If the task is rated as not interpretable at the previous stage, the attribution stage is skipped and the annotator proceeds to the next task in the queue.

Appendix G. Annotation User Interface for Table-to-Text Tasks

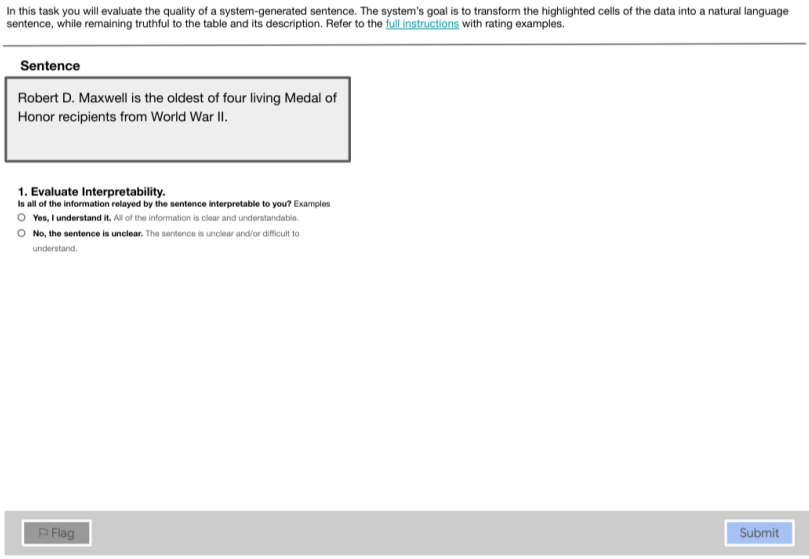


Figure G.10
Interpretability stage. The source table and its description are hidden. During training and the pilot the justification element is shown, but only if the task is rated as not interpretable.

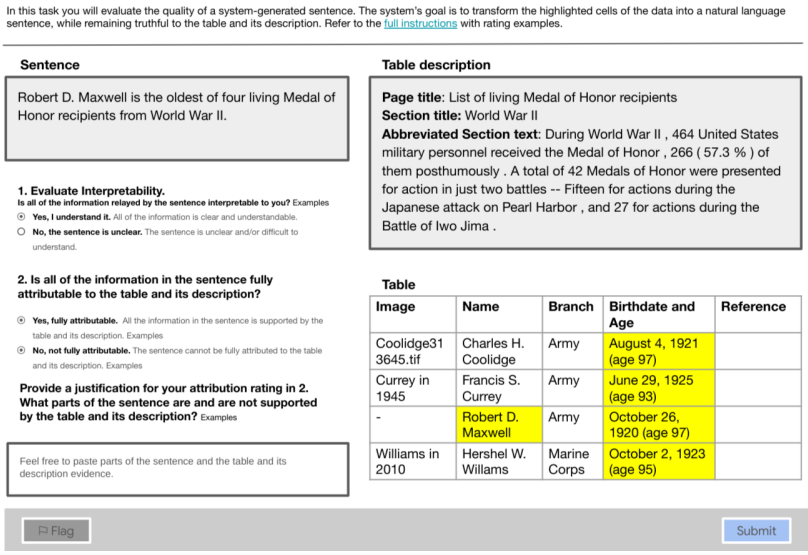


Figure G.11
Attribution stage. The source table and its description are shown. The rendering preserves highlighted cells from the ToTTo data. During training and the pilot the justification element is shown for all ratings in the second stage. If the task is rated as not interpretable at the previous stage, the attribution stage is skipped and the annotator proceeds to the next task in the queue.

Acknowledgments

The authors would like to thank anonymous reviewers for their insightful comments and suggestions. Additionally, we would like to thank Roee Aharoni, Sebastian Gehrmann, Mirella Lapata, Hongrae Lee, Shashi Narayan, Ankur Parikh, Fernando Pereira, and Idan Szpektor for their detailed feedback on the manuscript and throughout the project. We would also like to thank Ashwin Kakarla and his team for making the human evaluation study possible and Alejandra Molina and Kristen Olson for their guidance on the task interface design. We are thankful to the larger Google Research community for the many discussions during the course of this project.

References

Adiwardana, Daniel, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Anantha, Raviteja, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534. https://doi.org/10.18653/v1/2021.naacl-main.44

Belz, Anja, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194.

Carston, Robyn. 1988. Implicature, explicature, and truth-theoretic semantics. In Ruth Kempson, editor, *Mental Representations: The Interface Between Language and Reality*. Cambridge University Press, pages 155–181.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184. https://doi.org/10.18653/v1/D18-1241

Choi, Eunsol, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461. https://doi.org/10.1162/tacl_a_00377

Dalton, Jeffrey, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of SIGIR*, pages 1985–1988. https://doi.org/10.1145/3397271.3401206

Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Durmus, Esin, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

Dziri, Nouha, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083. https://doi.org/10.1162/tacl_a_00506

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault

Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120. https://doi.org/10.18653/v1/2021.gem-1.10

Gopnik, Alison and Henry M. Wellman. 1992. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2):145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Grice, Herbert P. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts*. Brill, Leiden, The Netherlands, pages 41–58. https://doi.org/10.1163/9789004368811_003

Gupta, Prakhar, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801. https://doi.org/10.18653/v1/2022.acl-long.263

Harrington, Leo A., Michael D. Morley, A. Šcedrov, and Stephen G. Simpson. 1985. *Harvey Friedman's Research on the Foundations of Mathematics*. Elsevier Science.

Honovich, Or, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870. https://doi.org/10.18653/v1/2021.emnlp-main.619

Howcroft, David M., Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Ladhak, Faisal, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? On mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421. https://doi.org/10.18653/v1/2022.acl-long.100

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Mehri, Shikib and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707. https://doi.org/10.18653/v1/2020.acl-main.64

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. https://doi.org/10.18653/v1/K16-1028

Nan, Feng, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894. https://doi.org/10.18653/v1/2021.acl-long.536

Nie, Yixin, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching

networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

Pagnoni, Artidoro, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829. https://doi.org/10.18653/v1/2021.naacl-main.383

Parikh, Ankur, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. https://doi.org/10.18653/v1/2020.emnlp-main.89

Pavlick, Ellie and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.

Rashkin, Hannah, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718. https://doi.org/10.18653/v1/2021.acl-long.58

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Santhanam, Sashank, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Z. Hakkani-Tür. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

See, Abigail, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. https://doi.org/10.18653/v1/P17-1099

Shuster, Kurt, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470. https://doi.org/10.18653/v1/2020.acl-main.222

Thorne, James, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs. *arXiv preprint arXiv:2104.00640*.

Thorne, James and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. https://doi.org/10.18653/v1/N18-1074, PubMed: 29492103

Wang, Alex, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020. https://doi.org/10.18653/v1/2020.acl-main.450

Welleck, Sean, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741. https://doi.org/10.18653/v1/P19-1363

Wilson, Deirdre and Dan Sperber. 2004. Relevance theory. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*. Blackwell, chapter 27, pages 607–632.

Wiseman, Sam, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2253–2263. https://doi.org/10.18653/v1/D17-1239

Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Zhong, Ming, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208. https://doi.org/10.18653/v1/2020.acl-main.552