SEMQA: Semi-Extractive Multi-Source Question Answering

Tal Schuster,[1]   Adam D. Lelkes,[1]   Haitian Sun,[2]   Jai Gupta,[1]

Jonathan Berant,[2]   William W. Cohen,[2]   Donald Metzler[1]

[1]Google Research   [2]Google DeepMind

Abstract

Recently proposed long-form question answering (QA) systems, supported by large language models (LLMs), have shown promising capabilities. Yet, attributing and verifying their generated abstractive answers can be difficult, and automatically evaluating their accuracy remains an ongoing challenge. In this work, we introduce a new QA task for answering multi-answer questions by summarizing multiple diverse sources in a semi-extractive fashion. Specifically, Semi-extractive Multi-source QA (SEMQA) requires models to output a comprehensive answer, while mixing factual quoted spans—copied verbatim from given input sources—and non-factual free-text connectors that glue these spans together into a single cohesive passage. This setting bridges the gap between the outputs of well-grounded but constrained extractive QA systems and more fluent but harder to attribute fully abstractive answers. Particularly, it enables a new mode for language models that leverages their advanced language generation capabilities, while also producing fine in-line attributions by-design that are easy to verify, interpret, and evaluate. To study this task, we create the first dataset of this kind, QuoteSum, with human-written semi-extractive answers to natural and generated questions, and define text-based evaluation metrics. Experimenting with several LLMs in various settings, we find this task to be surprisingly challenging, demonstrating the importance of QuoteSum for developing and studying such consolidation capabilities.[1]

1 Introduction

Large Language Models (LLMs) have recently demonstrated impressive capabilities across NLP tasks (Srivastava et al., 2022) and in particular, in answering general questions (OpenAI, 2023; Google et al., 2023). As a result, an increasing number of users interact with such models when seeking information. Although LLMs often return high-quality and correct results, they can still make convincing-sounding mistakes (e.g., relying on false or out-dated information (Dhingra et al., 2022; Schuster et al., 2020, 2021)). Added citations can help the reader verify the generated answers (Bohnet et al., 2023; Gao et al., 2023), but might be inaccurate themselves and add a false impression of reliability (Liu et al., 2023; Yue et al., 2023). Furthermore, evaluating fully-abstractive long-form QA systems remains a challenge, mostly involving either running another large QA model or expensive human ratings.

In this work, we introduce the task of semi-extractive multi-source QA (SEMQA) for answering multi-answer questions. Specifically, given a question and a set of retrieved passages, the goal is to generate a summarized and well-grounded answer that interleaves verbatim extracted spans of factual statements with free-text connectors.

Including explicitly extracted spans in long generated answers provides multiple advantages. For example, it readily supports fast and easy verification by the reader—only requiring to confirm the context in the source. Also, it simplifies and standardizes the evaluation process. Instead of relying on model-based metrics that are expensive to run, hard to interpret, and depend on the performance of the measuring model (Deutsch et al., 2022), extracted spans can be readily evaluated for both recall and precision with standard string-matching metrics (Rajpurkar et al., 2016). Finally, explicitly quoting spans can simplify the expensive auto-regressive generation process of LMs, enabling potential efficiency gains (Mallinson et al., 2022; Schuster et al., 2022).

Consider the example in Figure 1 that asks when was *"I'll be home for Christmas"* released. A fully-extractive format would likely only return the set of short-answers without any contextualization (e.g.,

---

[1]Dataset and code are available at: https://github.com/google-research-datasets/QuoteSum

1363

Question: *When was i'll be home for Christmas released?*

Sources:

[1] I'll Be Home for Christmas:
Songwriter and later producer and manager for The Platters, Buck Ram, who said he previously wrote a poem and song with the same title, was credited as a co-writer of the song following a lawsuit brought by Ram's publisher, Mills Music. Bing Crosby's original 1943 release of the song on Decca Records listed only Walter Kent and Kim Gannon as the songwriters on the record label. Later pressings added the name of Buck Ram to the songwriting credit.

[2] I'll Be Home for Christmas (EP):
I'll Be Home for Christmas is a Christmas-themed extended play released by Epic Records and Syco Entertainment on November 24, 2014. L.A. Reid served as the EP's executive producer, and it is Epic Records' first-ever holiday extended play.

[3] I'll Be Home for Christmas (1998 film):
I'll Be Home for Christmas is a 1998 American Christmas family comedy film directed by Arlene Sanford. The plot follows a college student who must make it from his campus in Los Angeles, California to his family's home in Larchmont, New York in time for Christmas dinner in order to win his father's Porsche. It stars Jonathan Taylor Thomas, Jessica Biel, Adam LaVorgna, Sean O'Bryan and Gary Cole and was released on November 13, 1998.

[4] I'll Be Home for Christmas (album):
It was released on October 7, 2008 through Razor & Tie. In his review for Allmusic, editor Anthony Tognazzini wrote that "one good holiday album deserves another, and McKnight's smooth, alluring voice sounds better than ever on 2008's "I'll Be Home for Christmas". The tunes are a bit more secular than "Bethlehem".

Quoted answer:

The song "I'll Be Home for Christmas" was originally released by Bing Crosby in 1943. The title would later be used in media like the film I'll Be Home for Christmas, which was released on November 13, 1998. I'll Be Home for Christmas is also a Christmas-themed extended play released by Epic Records and Syco Entertainment on November 24, 2014. Additionally, an album with the same title was released on October 7, 2008.

Figure 1: In SEMQA, the answer combines information from multiple sources while explicitly extracting factual spans (highlighted in different colors per source), and connecting them into a coherent well-grounded passage.

1943; November 24, 2014;...). Instead, in SEMQA, dates, relevant entity names, and factual statements are extracted and glued together into a cohesive and well-grounded passage.

To study the SEMQA task, we construct a dataset of questions, relevant passages, and human-written semi-extractive answers. Our dataset, named QuoteSum, includes both natural multi-source questions (Kwiatkowski et al., 2019; Min et al., 2020) and generated multi-answer questions (Lewis et al., 2021). We created an annotation pipeline where writers are asked to compile answers while using a tool that visualizes and keeps track of copied spans. In the task definition, we ask writers to use the provided quoting mechanism for all factual statements, while ensuring fluency of the text. Writers can choose to skip certain sources if they don't include additional helpful content and cite only a subset of the given sources, simulating possible noise in the retriever module (see §4).

We choose to focus on the multi-source setting since we argue it to more clearly benefit from this style. However, our approach can be easily relaxed to single-source and single-answer questions. By focusing on multi-answer questions we also emphasize the main challenges of this task, requiring models to (1) correctly find an answer in each source; (2) extract relevant context from each source to contextualize the short answers; and (3) effectively consolidate the different answers according to their relations with each other and to the question.

We use QuoteSum to fine-tune large QA models and evaluate both supervised and in-context LLMs. We find LLMs to obtain remarkable performance with only a few QuoteSum examples. However, QuoteSum-tuned models perform best, indicating

the usefulness of our data. We also run a user study comparing SEMQA answers to cited abstractive answers (Gao et al., 2023), and find SEMQA answers to be more comprehensive and much easier to be manually verified by readers.

Overall, our experiments reveal that the SEMQA task is surprisingly challenging even for modern LLMs. We define text-based metrics for SEMQA, creating easy-to-compute model-free measures to track the progress of models towards solving this task, capturing the fluency, preciseness and comprehensiveness of the generated quoted answers.

Our main contributions include:

- Introducing and formulating the task of semi-extractive multi-source QA (SEMQA).

- Creating QuoteSum, the first dataset for this task, including multi-answer questions with high-quality human-written semi-extractive answers.

- Evaluating different LLMs using both text-based metrics and via a user study, revealing SEMQA's challenges and facilitating future research.

2 Related Work

Several open-domain QA datasets use short extracted spans or entities as target answers (Berant et al., 2013; Joshi et al., 2017; Kwiatkowski et al., 2019; Rajpurkar et al., 2016, 2018). The recent advancements in LLMs allowed expanding automatic QA systems to more challenging settings such as long-form answers (Fan et al., 2019), answers with multiple entities (Amouyal et al., 2023; Zhong et al., 2022), and ambiguous questions (Min et al., 2020; Stelmakh et al., 2022). While most recent LLM-based QA models focus on abstractive free-text closed-book or retrieve-then-read set-

tings (Bohnet et al., 2023; Lewis et al., 2020b), evaluating these models remains a challenge (Min et al., 2023), even when answers have citations (Gao et al., 2023; Huang and Chang, 2023; Kamalloo et al., 2023; Li et al., 2023; Liu et al., 2023; Malaviya et al., 2023; Yue et al., 2023)—mostly involving expensive model-based measures (Honovich et al., 2022; Rashkin et al., 2022).

Similar to our motivation to improve the faithfulness of generated answers, Potluri et al. (2023) suggest an extract-and-decontextualize pipeline that starts with a fully extractive long-form answer, and edits it for improving fluency. However, they focus on single-source questions which don't require multi-answer consolidation.

Regarding evaluation, recent studies identified some limitations of lexical metrics for general QA evaluation (Bulian et al., 2022; Wang et al., 2023). However, our task formulation that limits the output space to the given sources, combined with the semi-extractive nature and collecting multiple references, allow us to define model-free evaluation metrics that overcome most of the discussed limitations (see §3). Ultimately, a practitioner might choose to combine multiple different evaluators.

3 Semi-Extractive Multi-Source QA

To formally define the SEMQA setting, we assume an input question q and a set of two or more passages $\mathcal{P}$ that possibly include an answer to question q. To simplify the setting, we assume here that the passages $p \in \mathcal{P}$ are already selected by some retriever that models $\mathbb{P}(p \mid q)$. Our goal is to generate an answer $(q, \mathcal{P}) \rightarrow a$ that (1) answers q, covering all components and aspects based on information from $\mathcal{P}$, (2) is as extractive as possible with respect to $\mathcal{P}$ (explicitly marking extractions and their source), and (3) is concise[2] and fluent.

There is a natural trade-off between the last two requirements for a. This balance is necessary since naïvely maximizing only (1) and (2) would lead to simply concatenating all passages in $\mathcal{P}$. In contrast, maximizing only (3) would lead to an empty answer. We resolve this tension by seeking to ensure extractiveness for entities and for core factual statements, and to prefer fluency otherwise.

By explicit quoting marks, we mean that each extracted span should be marked as such, with an indicator for which passage it was taken from. The

passage indicator is important for distinguishing identical spans across passages (e.g., *"I'll be home for Christmas"* in Figure 1) and connecting each statement in the answer with the original supporting source. For brevity, we refer to such explicitly marked semi-extractive answers as *quoted answers*, even though the answer is not required to include textual quotation marks around quoted spans.

We note that some questions might not be adequately answered with this format. Moreover, the feasibility of composing a sufficient answer also depends on the given passages. We argue (and demonstrate), though, that many questions can be answered in a semi-extractive format. This format provides several key advantages described below.

Verifiability. It is generally easier to verify answers that directly quotes factual statements compared to verifying answers that paraphrase information, even when sentences are attributed to supporting sources. In the latter case, the reader would need to decompose the generated text into statements, review the sources, locate relevant information pieces, and compare the generated statements against them. In contrast, when factual statements are extracted verbatim, verifying the generated text broadly simplifies to ensuring that the statements weren't taken out of context. The benefits of quoted spans are even more pronounced in sentences that aggregate information from multiple sources. Rather than citing many sources, leading to combinatorial growth, attributing at a span-level drastically simplifies the verification task. In §5.1, we empirically show that this hypothesis holds with human readers.

Automatic evaluation. Another advantage of constraining the output space to include extracted spans is enabling the use of well-defined string-based evaluation metrics for examining both precision and recall of generative models against expert-written references. Essentially, this formulation bypasses the ongoing challenge of evaluating free-text answers. Current solutions rely on model-based factual consistency measures (Bohnet et al., 2023; Honovich et al., 2022; Rashkin et al., 2022). However, it is hard to establish standardized evaluation protocols with such measures. For example, while larger backbone evaluation models often perform better, sometimes smaller models are used due to cost and speed considerations. Moreover, evaluating long generated outputs with multiple references exposes even further complexities, as it might re-

---

[2]Our "concisenss" objective is within the space of semi-extractive answers. Naturally, unconstrained fully-abstractive answers can be more concise, but don't satisfy objective (2).

quire decontextualization (Choi et al., 2021) and aggregating multiple noisy classifications (Laban et al., 2022). Finally, black-box model-based measures are hard to interpret at an instance-wise level. In contrast, we introduce text-matching metrics that support multiple sources and are simpler, faster, and easily intepretable.

Precision of generated text and attributions. Finally, beyond simplifying the verification and evaluation process, using extracted spans from reliable sources for generating factual claims could help prevent generation mistakes and model hallucinations, both in the text itself and in the text-attribution pairing. By enforcing the generative model to explicitly mark the boundaries and the source of each extracted span, the attribution to the correct source is provided by design.

3.1 Evaluation metrics

As discussed above, one of the advantages of SEMQA is in allowing pure string-based evaluation that is interpretable and fast to compute, and does not require additional annotations such as disambiguating questions. We formulate the following measures to evaluate answers for fluency, attribution preciseness, and comprehensiveness.

Fluency. Following the ASQA evaluation setting (Stelmakh et al., 2022), we use the ROUGE-L score (Lin, 2004) to compare generated answers with reference answers, after stripping any attribution marks. We take the maximum score across all human-written references.

Preciseness. To evaluate the quality of the extracted spans in the generated answer, we compute the normalized token-F1 score (Rajpurkar et al., 2016) for each source separately, and average across sources, taking the maximum score per source across reference quoted answers:

$$\text{Sem-F1}(a_i, \mathcal{A}_i) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \max_{\hat{a} \in \mathcal{A}_i} \left(\text{F1}(\psi_k(\hat{a}), \psi_k(a_i)) \right)$$

where $\mathcal{A}_i$ is the set of human-written reference answers for question q_i, K is the number of input sources, and $\psi_k(a)$ is a function that keeps only the tokens explicitly marked as extracted from source k in answer a. We use the F1 score to measure both the precision of spans (not extracting redundant tokens that don't answer the question), and their recall (extracting all answers and helpful context).

Comprehensiveness. We also measure short-answer recall to capture the aspect coverage of the generated quoted answer:

$$\text{Sem-Rec}(a_i, \mathcal{S}_i) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \max_{\hat{s} \in \mathcal{S}_{i,k}} \left(\text{Rec}(\hat{s}, \psi_k(a_i)) \right)$$

where $\mathcal{S}_i$ is the set of short answers,[3] by source, covered in each of the human-written references, and $\text{Rec}(\cdot, \cdot)$ is the token-level recall. A perfect score would mean that all short answers appearing in at least one of the references are also in the generated answer, and attributed to the right source.

Combined SEMQA score. The Fluency score ignores any attribution marks, while Preciseness only measures extraction (i.e., attribution) quality, ignoring the free text portions of the answer. Therefore, we compute their geometric mean, following ASQA (Stelmakh et al., 2022), to obtain a single score that reflects the overall answer quality:

$$\text{SEMQA} \coloneqq (\text{Sem-F1} \cdot \text{ROUGE-L})^{0.5}$$

4 The QuoteSum Dataset

To study the SEMQA task, we create a new multi-answer QA dataset with $(q, \mathcal{P}, a)$ triplets, where a is a human-written semi-extractive answer. We call this dataset QuoteSum, as it includes quoted answers to the input questions, based on the provided sources. We use square brackets and the index of the respective source to explicitly mark the extracted spans. For example, the first sentence of the answer in Figure 1 would be: `The song "[ 1 I'll Be Home for Christmas ]" was originally released by [ 1 Bing Crosby ] in [ 1 1943 ]` to mark the spans extracted from source 1. Throughout this paper, we replace the textual marks with colored highlights for presentation.

Next, in §4.1 we describe our process for collecting $(q, \mathcal{P})$ pairs based on the PAQ (Lewis et al., 2021) and NQ (Kwiatkowski et al., 2019) datasets. Then, in §4.2 we detail our crowd-sourcing task for writing quoted answers.

4.1 Collecting multi-answer questions

Since we focus on multi-answer questions, we first need to collect the questions and candidate answers that will be provided to the writers for producing the semi-extractive answers. We obtain both machine-generated questions from the PAQ dataset (Lewis et al., 2021) and human-written

---

[3] We use the provided short answers from the originating PAQ and AmbigQA examples.

1366

PAQ	Why was the porsche 911 rs built?
	Where did bob dylan tour in 1978?
	Which cities in india have police commissionerates?
NQ	Who is the actress that portrays wonder woman?
	When did rocky horror picture show come out?
	Who was british pm and viceroy during quit india movement?

Table 1: Example questions from QuoteSum dataset, requiring aggregating information from multiple sources.

questions from the AmbigQA (Min et al., 2020) subset of NQ (Kwiatkowski et al., 2019). Table 1 presents example selected questions.

Probably-Asked Questions (PAQ). Our first set of questions is based on the PAQ dataset (Lewis et al., 2021), a collection of 65 million probably-asked questions that were automatically generated from Wikipedia passages. For each passage, possible answer spans are extracted with a BERT (Devlin et al., 2019) model. Then, a question is generated with a BART-base (Lewis et al., 2020a) question generation model trained on NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and SQuAD (Rajpurkar et al., 2016).

We start with the large PAQ collection and look for multi-answer questions, meaning instances where the same question was independently generated from two different target answers and passages. Therefore, we merge all (q, p, s) triplets of question, passage, and short answer by identical questions to obtain a mapping $q \to [(p_1, s_1), \dots, (p_k, s_k)]$.

To avoid including multiple instances of the same answer, we apply a series of filters.[4] For each question q, we sort the (p_i, s_i) by $\mathbb{P}(s_i \mid q, p_i)$ according to a T5-XXL (Raffel et al., 2019) QA model trained on SQuAD, and keep only answers with score of at least 0.5. Then, we run over $(p_i, s_i)_{i=1}^k$ and filter out any instances where either (1) the Wikipedia page of p_i was already included; (2) the length of s_i is less than 4 words; (3) $s_i \in p_j$ for $j < i$; or (4) $\phi(s_i, s_j) = 1$ for $j < i$.

The function $\phi(x, y)$ is a binary answer similarity function that returns 1 if at least one of the following conditions are true: (1) the Levenstain distance $\text{Lev}(x, y) \leq 10$; (2) the word-based intersection-over-union $\text{IoU}(x, y) > 0.75$; or (3) the score of a semantic answer-similarity classifier $\text{BEM}(x, y) > 0.5$ (Bulian et al., 2022). The first term is useful for identifying different spellings

---

[4]The initial large size of this dataset allows us to filter aggressively as we can afford false positives.

of the same name (e.g., Céline and Celine). The second term helps for finding lists that are ordered differently. The last semantic BEM-based term catches other more subtle duplicates.

We also find that some slightly different questions are similar in meaning. To automatically merge similar questions, we vectorize them with TF-IDF and merge questions with cosine similarity greater than 0.9. When merging instances, we repeat the same answer filtering process above, and keep the question obtaining higher average score against the answers, by the same T5 QA model.

This process resulted in about $170K$ questions, each with at least two answer passages. At first, the questions were highly unbalanced in type (e.g., many *what* and *who* questions) and in the number of answers. Therefore, we partition the data by questions starting with *what, who, where, when, how, which, why*, questions including *stand for* or "other". We also partition by the number of answers. Then, we perform balanced random sampling when selecting the examples for QuoteSum.

Natural Questions (NQ). We also collect human-written questions based on the NQ dataset (Kwiatkowski et al., 2019). Since we are interested in multi-answer questions, we use the subset of questions identified by Min et al. (2020) to be multi-answer, together with their annotated short answers s_i to disambiguating questions. We also collect the Wikipedia passages p_i that contain s_i. We notice that some disambiguating questions q_i' could be slightly nuanced. To collect passages that directly answer the original questions q, we filter out (p_i, s_i) pairs for question q where $\mathbb{P}(s_i \mid q, p_i) < 0.5$ according to a T5-XXL SQuAD-trained model.

Since some short answers to the same question are coming from the same passage, we remove passages that have word intersection ratio greater than 0.4 with another passage. Finally, we keep only questions with at least two different answers.

Similar to the PAQ set, we see an imbalance in the question types (with many *who* and *when* questions). Again, we balance the sampling across question types and number of answers to diversify the examples to include in QuoteSum.

4.2 Writing quoted answers

We now detail our annotation task for creating human-written semi-extractive answers a for our gathered questions and passages $(q, \mathcal{P})$. We build

a web interface for writers that presents a question and a list of passages (see Figure A.1). Writers can select spans from input passages and copy them to the answer text box. These copied spans are colored to match the respective input source to help the writer keep track of the covered sources. Writers can also use free text to connect the extracted spans and complete the answer. The submitted answers include our special marks for signaling which spans were copied, and their originating passage.

Writers are asked to create a concise summary that answers the question while copying informative spans from the sources whenever possible, and using free text with a diverse set of connectors. They are told to cover as much information as possible, as long as it directly relates to the given question. See Appendix A for more details.

We only provide the writers with the questions and matching passages, without revealing any short answer spans. We choose so to avoid biasing the writers to follow any noisy answers that might have been included through the scraping process. Therefore, writers can choose not to use all input passages if they find some to be irrelevant. We view this selection process as part of the task, and keep all input passages in our dataset, realistically simulating a slightly noisy retrieval output provided to the answer generation model.

4.3 Dataset statistics

QuoteSum includes 4,009 semi-extractive answers to 1,376 unique questions (984 from PAQ and 392 from NQ). As reported in Figure 2 and illustrated in Table 1, QuoteSum includes a diverse set of multi-answer questions. See App. C for more details.

5 Experimental Evaluation

We study the performance of both fine-tuned LMs, and in-context few-shot LLMs, as described below. We follow the evaluation metrics described in §3.1, and also run a user study as detailed in §5.1.

Fine-tuned models. We fine-tune different sizes of T5 models (Raffel et al., 2019) on the QuoteSum training set. We also fine-tune FLAN instruction-tuned T5 models (Wei et al., 2022). See Appendix D for more details.

In-context learning. We evaluate variants of few-shot PaLM2 (Google et al., 2023). We construct a dynamic prompt that collects examples for each test question q_i by retrieving the most similar
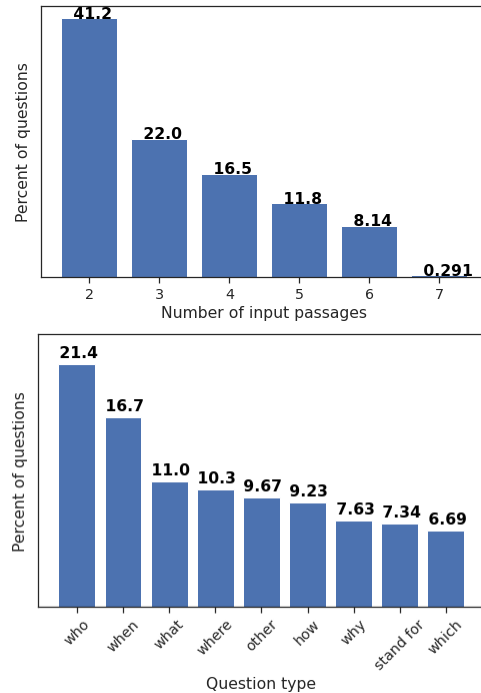
Figure 2: Diversity of questions in QuoteSum.

questions from the QuoteSum training set according to Sentence-T5 (Ni et al., 2022) embedding cosine similarity (QSum). We also experiment with the ALCE (Gao et al., 2023) prompt that uses fully-abstractive answers with in-line citations (e.g., [1] to cite source 1), and with Qsum-S prompts that convert QSum examples to sentence-level citations. See Appendix D for details and example prompts.

5.1 Results and analysis

Table 2 summarizes the results of the fine-tuned models and the in-context learners. Fine-tuned models generally perform better at generating quoted answers as they increase in size, both in terms of short-answer coverage (Sem-Rec), and in terms of fluency (Rouge-L) and preciseness of the extracted spans (Sem-F1). Also, the Flan-T5 model family is significantly better than the original T5, possibly due to the set of instruction-following tasks they were trained on that might have enabled better adaptation to new output formats. For example, Flan-T5 base is matching the scores of T5 XL that has about $\times 13$ more parameters.

The few-shot learners obtain relatively high performance with only a few examples, demonstrating their sample efficiency. However, they still lack behind the fine-tuned models, suggesting the value of our quoted summaries for training models to generate consolidated quoted answers.

Interpreting the string-based metrics. Our metrics can help understand the strengths and weaknesses of different models. For example, 5-shot models obtain reasonably high Sem-Rec score, with PaLM2 Unicron achieving as high as 88.87, meaning that answers successfully cover multiple aspects of the question. The lower Sem-F1 scores show that the model can be more precise in extracting only relevant spans for answering the question.

Also, comparing the few-shot models to the T5 large model, we see that the Rouge-L scores are comparable while the Sem-F1 score of the T5 model is much higher. This suggests that the few-shot model matched the fluency of the T5 model, but didn't mark all extracted spans as quotes.

This is demonstrated in Table 4 with an example answer by PaLM2 Bison 4-shot. The quoted answer comprehensively extracts all short-answers from the given passages while also helpfully contextualizing each option. While the extracted spans provide useful attributions to the supporting passages, the preciseness of this output can be improved. For example, "Susan Egan" can be dropped, and the year of each show that is given in free-text should instead be quoted from the respective passage, like in the human-written reference.

String-based vs. model-based metrics. For comparison, we also report the model-based evaluation metrics from ASQA (Stelmakh et al., 2022). The Disambig-F1 metric requires running an extractive RoBERTa large (Liu et al., 2019) QA model on a list of disambiguating questions and the generated long answer, and comparing each predicted answer span with the gold short-answer. Since we don't have disambiguating questions for PAQ examples, we only evaluate the portion of our test set originating from AmbigQA. As Table 3 shows, our Sem-F1 text-based measure ranks the examined models the same as Disambig-F1, without relying on any QA models or disambiguating questions.

Comparing to answers with citations. One alternative approach for attributing answers is to generate the answer text with citations. We evaluate SEMQA against that approach using both automatic metrics and human ratings. To this end, we compare against the recently proposed few-shot ALCE method (Gao et al., 2023) that, similar to us, prompts an LLM with the question and a set of pre-retrieved passages. We use the ALCE ASQA

Model	Rouge-L	Sem-F1	Sem-Rec	SEMQA
T5 small	54.29	57.66	68.30	55.95
T5 base	60.39	67.69	78.04	63.94
T5 large	63.90	71.55	83.68	67.61
T5 XL	65.36	73.05	84.88	69.10
T5 XXL	70.49	80.60	91.25	75.38
Flan-T5 small	58.33	63.05	78.24	60.65
Flan-T5 base	65.08	73.55	86.32	69.19
Flan-T5 large	71.59	81.62	91.79	76.44
Flan-T5 XL	72.89	83.82	94.38	78.17
Flan-T5 XXL	73.36	84.20	93.68	**78.59**
PaLM2 Bison 1-shot	60.04	50.11	52.62	54.85
PaLM2 Bison 4-shot	62.71	62.11	77.32	62.41
PaLM2 Bison 5-shot	63.45	65.40	80.09	64.42
PaLM2 Unicorn 5-s.	64.13	68.28	88.87	66.18

Table 2: Test results of fine-tuned and few-shot models.

Ver.	n	Prompt	Rouge	Sem-F1	SEMQA	Dis.-F1	DR
Bis.	4	QSum	64.46	58.04	61.17	45.32	54.05
Bis.	5	QSum	61.75	59.15	60.44	46.06	53.33
Uni.	5	QSum	64.23	63.76	64.00	**49.05**	**56.13**
Bis.	4	QSum-S	60.17	–	–	39.03	48.46
Bis.	4	ALCE	54.73	–	–	34.05	43.17

Table 3: PaLM2 n-shot results on the AmbigQA (Min et al., 2020) subset of QuoteSum test set using our QuoteSum prompt (QSum), ALCE prompt (Gao et al., 2023), and QuoteSum prompt converted to sentence-level citations (QSum-S). We also report the ASQA RoBERTa-based Disambig-F1 score and the combined DR score (Stelmakh et al., 2022).

light instruction prompt[5] that includes 4 examples of fully-abstractive answers with in-line citations to input sources.

Table 3 reports the ASQA metrics (Stelmakh et al., 2022) for PaLM2 with our QuoteSum prompt and the ALCE prompt. Since in ALCE spans are not explicitly quoted, even if copied from the input passages, they cannot be scored with our semi-extractive metrics. The QuoteSum prompt (QSum) leads to higher scores, both in ROUGE-L and in Disambig-F1, suggesting that beyond better interpretability, our semi-extractive approach also provides more precise answers. To more directly compare the two approaches, we automatically (using a regex function) convert our QuoteSum examples to remove span-level attributions and instead add respective sentence level citations (QSum-S). Then, instead of the ALCE 4 examples, we dynamically retrieve examples from the training set by question similarity (like in QSum). While improving over the ALCE prompt, the converted format (QSum-S) still scores lower than the semi-extractive format (QSum), indicating the benefits of explicitly extracting spans.

---

[5] https://github.com/princeton-nlp/ALCE

Question	*Who plays as beast in beauty and the beast?*
Sources	[1] Beauty and the Beast (1987 TV series): Beauty and the Beast is an American fantasy-drama television series which first aired on CBS from September 25, 1987 to August 4, 1990. Creator Ron Koslow's updated version of the fairy tale has a double focus: the relationship between Vincent (Ron Perlman), a mythic, noble man-beast, and Catherine (Linda Hamilton), a [...] [2] Beauty and the Beast (musical): After completing tryouts in Houston, "Beauty and the Beast" premiered on Broadway on April 18, 1994 , starring Susan Egan and Terrence Mann as the eponymous Belle and Beast, respectively. The musical opened to mixed reviews from theatre critics, but was a massive commercial success [...] [3] Beauty and the Beast (1991 film): "Beauty and the Beast" focuses on the relationship between the Beast (voice of Robby Benson), a prince who is magically transformed into a monster and his servants into household objects as punishment for his arrogance, and Belle (voice of Paige O'Hara), a young woman whom he imprisons in his castle to become a [...] [4] Beauty & the Beast (2012 TV series): Kristin Kreuk and Jay Ryan star in the title roles alongside Austin Basis, Nina Lisandrello, Nicole Gale Anderson, Sendhil Ramamurthy, Max Brown, Brian J. White, Amber Skye Noyes, and [...] [5] Beast (Beauty and the Beast): In all animated film appearances, the Beast is voiced by American actor Robby Benson . The 1991 animated film was adapted into a Broadway musical in 1994, with the role being originated by American actor Terrence Mann. Dan Stevens portrays a live-action version of the character in the 2017 live-action adaptation of the original 1991 film. [6] Beauty and the Beast (2014 film): On the way home through the forest, the merchant loses his way, his horse slips and is injured, and they are attacked by wolves. He laments that he has "not even a weapon to finish off" the poor horse. The merchant stumbles upon the magical domain of the Beast (Vincent Cassel). [7] Beauty and the Beast (1946 film): Awakened by a loud roar, he wanders the castle's grounds. Remembering that Belle asked for a rose, he plucks a rose from a tree which makes the Beast (Jean Marais) appear. The Beast threatens to kill him for theft but suggests that one of his daughters can take his place.
SEMQA (human-written)	The actors who play the Beast in the various versions of Beauty and the Beast include Dan Stevens in the 2017 live-action adaptation ; Vincent Cassel in the 2014 film ; Jay Ryan in the 2012 TV series ; Terrence Mann in the musical that premiered on Broadway on April 18, 1994 ; Robby Benson in The 1991 animated film ; Ron Perlman in the 1987 TV series and Jean Marais in the 1946 film .
SEMQA (model)	Ron Perlman plays as Beast in Beauty and the Beast (1987 TV series). Susan Egan and Terrence Mann play as Belle and Beast respectively in Beauty and the Beast (musical). Robby Benson is the voice of Beast in Beauty and the Beast (1991 film). Kristin Kreuk and Jay Ryan star in the title roles in Beauty & the Beast (2012 TV series). Dan Stevens portrays a live-action version of the character in the 2017 live-action adaptation of the original 1991 film. Vincent Cassel plays as Beast in Beauty and the Beast (2014 film). Jean Marais plays as Beast in Beauty and the Beast (19 [...]
ALCE (model)	The Beast is played by Ron Perlman in the 1987 TV series [1], Terrence Mann in the 1994 musical [2], Robby Benson in the 1991 film [3], Jay Ryan in the 2012 TV series [4], Vincent Cassel in the 2014 film [6], and Jean Marais in the 1946 film [7].

Table 4: Human-written example answer from QuoteSum, and generated answers of 4-shot PaLM2 Bison with our SEMQA prompt, and with the ALCE prompt. Colored spans show the spans that the writers or the model marked as explicit extractions from the respective source, by color. The extracted spans for the human-written answer are also highlighted in the input sources.

We also run a user study where human raters (a different pool than the one creating QuoteSum) were asked to score generated answers for fluency, comprehensiveness, correctness, and attribution helpfulness, using a 5-level Likert scale (See Appendix B for details).[6] We compare the generated answers of PaLM2 Bison 4-shot prompted with either SEMQA or ALCE examples, and SEMQA human-written answers from QuoteSum. As Figure 3 shows, ratings were overall high for both models, yet still below human performance. The SEMQA model scored slightly lower than ALCE on fluency, indicating the challenge in "glueing" the extracted spans in a coherent way. However, the human-written references obtained virtually perfect fluency score, demonstrating the potential of semi-extractive fluent answers.

The SEMQA model performed better on comprehensiveness, and significantly better on correctness, and attribution helpfulness. This supports our hypothesis that semi-extractive generation reduces hallucinations, and that span-level attributions help the reader verify the generated text.

Human verification with SEMQA is easier. We also measure the rating time for each example. Since SEMQA answers were generally longer than ALCE answers, we normalize by the number of words in each answer (excluding any attribution marks) to compute the average seconds per word. As shown in Figure 4, rating generated SEMQA answers was more than twice faster than rating ALCE abstractive answers. This suggests that in addition to being higher-quality answers (according to ratings), SEMQA answers with clearly marked quotes are also easier to evaluate and verify by readers.

---

[6] We choose scoring over side-by-side evaluation due to the clear guidelines and factual nature (Bansal et al., 2024).
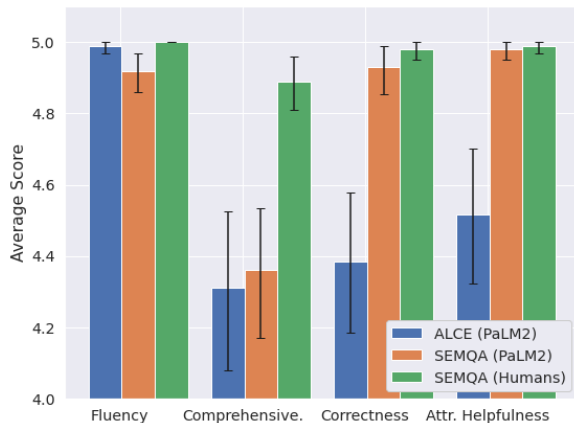
Figure 3: Average human ratings (1-5 scale) for answers of PaLM2 Bison 4-shot prompted with either SEMQA or ALCE examples, and for human-written SEMQA answers from QuoteSum. Error bars show 95% confidence intervals obtained with bootstrapping.
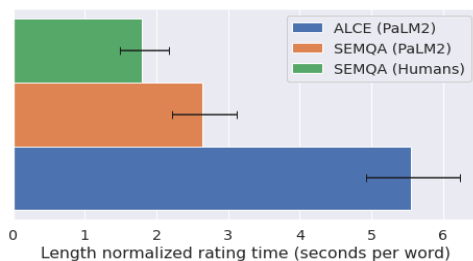

Figure 4: Manually evaluating and verifying generated SEMQA answers is twice faster than evaluating fully abstractive answers. Error bars show the 95% confidence intervals obtained with bootstrapping.

Solving SEMQA is non-trivial. The results of our experiments and user-study show that both fine-tuned and in-context LLMs can achieve noteworthy performance on SEMQA, thanks to our QuoteSum dataset and task formulation. However, the results also show that current models still lack behind human performance, indicating the challenges of this task. To further validate the complexity of this task and dataset, we also verify that trivial solutions cannot solve it or trick our string-based metrics.

Previous studies have found naïve baselines such as taking the first few sentences of the source document to perform well on certain extractive summarization datasets (Grenander et al., 2019; Zhong et al., 2019). Table 5 reports the QuoteSum performance of a baseline that extracts the first or last k sentences from each source and simply concatenates them as the answer. As expected, these baselines perform poorly (compared to all models; see Table 2). Unsurprisingly the Sem-Rec score, that

Baseline	Rouge	Sem-F1	Sem-Rec	SEMQA
Lead 1	29.58	30.01	39.99	29.79
Lead 2	34.32	32.74	60.36	33.52
Lead 3	33.69	34.96	77.44	34.32
Lead 4	31.52	33.45	88.55	32.47
Lead 5	29.87	31.96	93.55	30.90
Tail 1	23.79	18.47	41.22	20.96
Tail 2	28.26	27.14	61.10	27.69
Tail 3	28.63	29.35	79.97	28.99
Tail 4	28.61	30.06	88.80	29.32
Tail 5	28.39	30.19	94.56	29.27

Table 5: Trivial baselines that simply extract the first or last k sentences from each source perform poorly.

only measures recall of short answers, increases with more sentences included. However, the Rouge-L, Sem-F1, and the aggregated SEMQA scores are very low as they also measure the precision of the selected spans and the fluency of the answer compared to the references.

6   Conclusion

We introduce the task of semi-extractive multi-source QA (SEMQA), and a new comprehensive dataset with human-written SEMQA examples (QuoteSum) with text-based evaluation metrics. Our experiments and analysis reveal the benefits of SEMQA for enabling LLMs to generate well-grounded answers that are fast to verify against external sources, and easy to interpret and evaluate against references—avoiding the reliance on model-based measures. We hope that SEMQA and the released data will facilitate further research on improving LLM capabilities around reliable QA.

We make the QuoteSum dataset and implementation of the SEMQA metrics available at: https://github.com/google-research-datasets/QuoteSum.

Limitations

Our study is limited in scope to English questions and answers, and to English Wikipedia articles as supporting sources. While we applied diversified balanced sampling for choosing multi-answer questions out of the larger NQ and PAQ collections, post our filtering and merging functions, the scope of our examined questions is still limited to the ones contained in those collections. Also, as we discuss throughout the paper, we don't aim to suggest that our SEMQA approach is ideal for all types of questions and answers. Rather, we argue that it is suitable for many multi-answer questions, and that it provides many benefits when appropriately followed.

In our dataset and experiments, we provide a set of retrieved sources that we collect following the process described in section 4. This setup is useful for objective evaluation of models, focusing on the reading comprehension and information consolidation capabilities. Many other notable QA datasets such as SQuAD (Rajpurkar et al., 2016) and Natural Questions (Kwiatkowski et al., 2019) also first focused on such setups, leaving extensions to open-domain systems to later work (Chen et al., 2017). One limitation of the closed-domain setup is that models might fail properly generalize to inputs from noisy retriever systems. However, as we discuss in the paper, our data collection process intentionally kept some noisy sources in the input which were not cited by the writers in the reference answers. Therefore, to perform well on QuoteSum, models should already learn to filter out irrelevant sources. Moreover, it is generally easier to further augment the dataset automatically with unused sources.

We also clarify that while the semi-extractive format might increase alignment with the supporting attributed sources, it is not bullet-proof against model hallucinations such as incorrect coreferences or other out-of-context issues (Zhang et al., 2023). We find semi-extractive answers to be easier to manually verify against the attributed source. Finally, it is important to note that the correctness of the generated answer also depends on the accuracy of the provided sources, which might include incorrect or misleading information. Here, to simplify the setting and keep it focused, we assume that the sources are provided by some possibly noisy retriever. We leave further studies on the retrieval part to future studies.

Broader Impact

We publicly release the dataset collected for this study, including the human-written semi-extractive answers. We hope that this will facilitate future research on enabling semi-extractive QA systems, further exploring their tradeoffs, and identifying the ideal question coverage for such models.

References

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

Hritik Bansal, John Dang, and Aditya Grover. 2024. Peering through preferences: Unraveling feedback acquisition for aligning large language models. In *The Twelfth International Conference on Learning Representations*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Eunsol Choi, 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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.

Tianyu Gao, Ho-Ching Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *ArXiv*, abs/2305.14627.

Rohan Anil Google, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *ArXiv*, abs/2307.02185.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv:2307.16883*.

Tom Kwiatkowski, 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.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-

based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Xinze Li, Yixin Cao2, Liangming Pan, Yubo Ma, and Aixin Sun. 2023. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. *ArXiv*, abs/2310.05634.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *ArXiv*, abs/2309.07852.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv*, abs/2305.14251.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Abhilash Potluri, Fangyuan Xu, and Eunsol Choi. 2023. Concise answers to complex questions: Summarization of long-form answers.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2022. Measuring attribution in natural language generation models.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan

Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-QA evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Xiang Yue, Boshi Wang, Kai Zhang, Zi-Yuan Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *ArXiv*, abs/2305.06311.

Shiyue Zhang, David Wan, and Mohit Bansal. 2023. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89, Hong Kong, China. Association for Computational Linguistics.

Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. 2022. Romqa: A benchmark for robust, multi-evidence, multi-answer question answering.

A Details of Answer Writing Task

We crowd-sourced native English speakers with Bachelor's degrees in Journalism or related fields to compile the semi-extractive answers as described in §4.2. A screenshot of the writing plugin used for this task is shown in Figure A.1. We constricted the answer to be no more than 100 words in order to keep it concise and query-focused. We find this limit to be sufficient for properly answering the examined questions.

Before starting the main writing task, we performed three pilot runs in which writers worked on 10 example questions. After each pilot run, we examined the written answers, provided feedback, refined the instructions, and also provided further reference example answers that the writers could learn from. Writers were also given the option to provide feedback to specific questions, which allowed us to discuss and clarify any confusions regarding the task.

The main writing task was performed in batches. After each batch, we randomly examine a subset of the answers and provide additional feedback and clarifications when needed. For instance, after the first batch, we identified that one of the writers produces overly-extractive answers that are not fluent. We sent these answers back for rewriting, and clarified the guidelines accordingly.

B Details of the User Study

In the generated answer rating task, raters were given a single example each time that includes the question, the sources, and the machine-generated answer. We rendered the outputs to highlight the extracted spans in SEMQA answers with colors that match the respective spans in the input sources (similar to the visualization demonstrated in Figure 1). For consistency across the two formats, we also post-process the SEMQA outputs (with a regular expression function) to add sentence-level citations similar to ALCE. Raters were only told

Figure A.1: A screenshot of our tool and short instructions for writing semi-extractive answers.

that attributions are provided as citations, and sometimes also as highlighted colored spans that match respective spans in the sources. All examples were shuffled and raters were not told they can originate from different models.

We used a different pool of raters than the one used for the QuoteSum dataset. We asked raters to mark their agreement with the following statements, using a five-level Likert scale:

1. **Fluency**: The machine-generated answer is fluent and coherent.
2. **Comprehensiveness**: The machine-generated answer is comprehensive in covering different aspects of the question.
3. **Correctness**: All information in the machine-generated answer is fully supported by the provided respective attributions.
4. **Attribution helpfulness**: The attributions in the machine-generated answer are helpful for evaluating its correctness.

We collected ratings for 100 outputs of each model with 3 repetitions by independent raters.

Inter-rater agreement was high (0.75 Krippendorff's alpha).

A screenshot of the rating plugin used for human evaluation of machine-generated answers is shown in Figure B.1.

C Additional Dataset statistics

As mentioned in Section 4.3, QuoteSum includes a diverse set of multi-answer questions with different question types. As Figure 2 shows, the examples also vary in the number of provided input passages. The maximum number of input passages in the dataset is seven, and most examples have two or three input passages.

We split the data to train, validation, and test sets with ratios 60%, 7%, 33%, respectively.[7] We ensure that the set of Wikipedia pages that the test set answers are based on are disjoint from the train and validation sets.

---

[7]We use a training split to enable fair comparison between finetuned and in-context learning (ICL) models. However, studies that only focus on ICL models might choose to use most or all of the collected dataset for evaluation.

Task

In this task, you are presented with a question and a machine-generated answer with attributions to supporting sources.

Please rate your agreement with the following statements:
1. **Fluency**: The machine-generated answer is fluent and coherent.
2. **Comprehensiveness**: The machine-generated answer is comprehensive in covering different aspects of the question.
3. **Correctness**: All information in the machine-generated answer is fully supported by the provided respective attributions.
4. **Attribution helpfulness**: The attributions in the machine-generated answer are helpful for evaluating its correctness.

Note:
- Attributions are provided as citations at the end of sentences to indicate the respective supporting sources for that sentence (for example, [1] [2] for sources 1 and 2), and sometimes also as colored highlighted spans that match respective spans (by color) in the related source.
- When replying, assume that the provided supporting sources are true and can be relied upon. You are evaluating the machine-generated answer, not the content of the supporting sources.

Question

who were the original judges on x factor?

Machine Generated Answer

The original judging panel of The X Factor (UK TV series) consisted of Louis Walsh, Sharon Osbourne and Cowell [1]. The original judging panel line-up of X Factor (Romanian TV series) in 2011 consisted of Adrian Sînă, Paula Seling, and Mihai Morar [2]. The original judging panel of X Factor (Danish TV series) consisted of Thomas Blachman, Lina Rafn and Remee [3]. The original judging panel line-up of The Pussycat Dolls in 2011 consisted of Simon Cowell, Paula Abdul and L.A. Reid [4]. The judges of X Factor (Icelandic TV series) were the talent agent and businessman Einar Bárðarson, rock musician Elínborg Halldórsdóttir and pop singer Paul Oscar [5]. The original judging panel line-up of The X Factor (Australian TV series) in 2005 consisted of Mark Holden, Kate Ceberano, and John Reid [6].

Supporting sources

[1] The X Factor (UK TV series) : onwards. The original judging panel consisted of Louis Walsh, Sharon Osbourne and Cowell. In 2005, Paula Abdul joined the show as a guest judge whilst Osbourne was away then joined the panel in 2006 for three sets of auditions. Brian Friedman briefly replaced Walsh in the fourth series, which also saw Dannii Minogue join the panel. Friedman left during the auditions, and Walsh replaced Friedman. Cheryl Cole replaced Osbourne in the fifth series. Gary Barlow, Kelly Rowland and Tulisa joined the panel in the eighth series as replacements for Cowell, Minogue and Cole. Rowland left before the ninth series and

[2] X Factor (Romanian TV series) : compete to ensure that their act wins the competition, thus making them the winning judge. The original judging panel line-up in 2011 consisted of Adrian Sînă, Paula Seling, and Mihai Morar. In May 2012, it was announced that Morar, Sînă, and Seling had all left and would not be returning for the second season. On 15 June, before auditions began, Dan Bittman, Delia Matache and Cheloo were confirmed as the three new judges. When the show was revived in 2014, Bittman and Cheloo were replaced by Horia Brenciu and Ștefan Bănică, Jr. On July 12, 2016 it was announced that

[3] X Factor (Danish TV series) : X Factor (Danish TV series) X Factor is the Danish version of "The X Factor", a show originating in the United Kingdom and is created by talent show judge and record and TV producer Simon Cowell. There has been eleven winners to date: Martin Hoberg Hedegaard, Linda Andrews, Thomas Ring Petersen, Sarah Skaalum Jørgensen, Ida Østergaard Madsen, Chresten Falck Damborg, Anthony Jasmin, Emilie Esther, Embrace, Morten Nørgaard and Place on Earth. The original judging panel consisted of Thomas Blachman, Lina Rafn and Remee. Soulshock and Pernille Rosendahl took the places of Blachman and Rafn in season 3. Blachman returned to

[4] The Pussycat Dolls : In May 2011, Scherzinger was hired as a judge on the first season of the American version of "The X Factor", alongside Simon Cowell, Paula Abdul and L.A. Reid. After her stint in the United States, Scherzinger joined "The X Factor" UK for its ninth series as the fourth permanent judge. In May 2013, Scherzinger was confirmed as a returning judge for the tenth series of "The X Factor" UK. In January 2014, it was reported that Scherzinger left Interscope Records and signed a multimillion-dollar recording deal with Sony Records. Scherzinger's album, titled "Big Fat Lie", was released in October 2014. Jessica Sutta released her debut album "I Say Yes" on March 3, 2017. With only two studio

[5] X Factor (Icelandic TV series) : X Factor (Icelandic TV series) X Factor was the Icelandic version of the popular British television show The X Factor. It premiered in Reykjavík on November 17, 2006 with actress Halla Vilhjálmsdóttir hosting the show. The judges were the talent agent and businessman Einar Bárðarson, rock musician Elínborg Halldórsdóttir and pop singer Páll Óskar had been judges in "Idol stjörnuleit", the Icelandic version of "Pop Idol". In the finals of "X Factor" held in Vetrargarðinum, Smáralind, a large shopping center in Iceland, singer and hair dresser Jógvan Hansen from Faroe Islands was the winner.

[6] The X Factor (Australian TV series) : broadcast on the Seven Network in Australia and on TV3 in New Zealand. The program was cancelled after its eighth season in 2016. The original judging panel line-up in 2005 consisted of Mark Holden, Kate Ceberano, and John Reid. When the show was revived in 2010, the judging panel was replaced by Natalie Imbruglia, Ronan Keating, Kyle Sandilands and Guy Sebastian. Imbruglia and Sandilands did not return for season three and were replaced by Natalie Bassingthwaighte and Mel B. Sebastian returned to the

Fluency

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |

Comprehensiveness

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |

Figure B.1: A screenshot of our tool for rating the quality of generated answers against the supporting sources.

D Implementation Details

To fine-tune the T5 models (Raffel et al., 2019) we use the T5X framework (Roberts et al., 2022). We use the T51.1 and Flan-T5 checkpoints and train the models on TPUv4 chips, other than the "large" or smaller Flan-T5 models that used TPUv3 chips. We trained the models for up to 25k steps with a batch size of 32 and learning rate of 1e-3 with the Adafactor optimizer (Shazeer and Stern, 2018). For reference, training Flan-T5 base on 4 TPUv3 took 1.5 hours. For each model, we picked the best checkpoint according to the ROUGE-L score on the validation set. We did not perform any hyperparameter tuning, and only retrain the T5-XL once with a new random seed, since the first training run of that model led to low training and validation scores.

Our few-shot evaluations are querying variants of the PaLM2 model (Google et al., 2023), which is available via API calls.[8] We use temperature 0.0 to increase factuality. An example of our QuoteSum prompt with the preceding instructions and two retrieved demonstrations is provided in Table D.2. We also include the respective QSum-S prompt

in Table D.3, which includes QuoteSum examples that were automatically converted to have ALCE-style sentence-level citations instead of span-level quotes, and has the light ALCE (Gao et al., 2023) prompt instructions and formatting.

D.1 In-context exemplar selection ablation

We evaluate the effect of our Sentece-T5 nearest neighbor approach for selecting in-context examples. Table D.1 shows that this approach provides better results compared to randomly selecting examples from the dataset. With random examples, the "fluency" (Rouge-L) drops a bit, and the "preciseness" and "comprehensiveness" scores drop more (by about 5 points each). This indicates that the random examples still help the model to generate relatively good answers, but the related questions are more helpful in guiding the model to extract the right spans from the sources. Interestingly the "random" 4-shot results are still higher than the "nearest neighbor" 1-shot (especially in comprehensiveness), suggesting that more Quote-Sum examples help the model align better with the SEMQA task, and that examples with more similar questions help the model extract better spans for the semi-extractive answer.

Overall, given the diversity of questions in

---

[8] https://developers.generativeai.google/products/palm

1377

QuoteSum, it is not surprising that different questions lead to different strategies for composing semi-extractive answers. Retrieving similar questions and their answers can help guide the model towards answering the examined question. This shows the value of our curated QuoteSum dataset.

Exemplars	Rouge-L	Sem-F1	Sem-Rec	SEMQA
Random	61.23	57.14	72.19	59.15
KNN	62.71	62.11	77.32	62.41

Table D.1: Our nearest neighbor question selection from the QuoteSum training set for the in-context few-shot prompt (§5) leads to better performance than with randomly sampled examples. Reported results are with 4-shot PaLM2-Bison model.

E   Additional Results

Table E.1 shows an example answer of PaLM2 Unicorn 5-shot model with a SEMQA QuoteSum-based prompt, and a PaLM2 Bison with ALCE prompt answer to the same question.

Answer the question by summarizing the given sources while explicitly copying spans from the sources. When copying a span, use brackets and the respective source number to indicate that this span was copied. Use explicit copying as much as possible and for all factual statements, while preserving fluency. Make sure to use all relevant sources and properly quote them. Here are some examples:

Question: how much power does a wind turbine produce?

[1] Compact wind acceleration turbine: It is generally thought that since the amount of power produced by a wind turbine is proportional to the cube of the wind speed, any acceleration benefit is potentially statistically significant in the economics of wind. As noted though this is an inaccurate as it ignores the impact of the exit to area ratio and is therefore an apples to oranges comparison. In the case of a typical CWAT/DAWT the power result in perfect theoretical operation once adjusted for the area of the shroud is actually the square of the velocity at the rotor. As the CWAT/DAWT diverges from theoretical function the power increase drops significantly according

[2] Sustainable architecture: roof ledge. Small-scale rooftop wind turbines have been known to be able to generate power from 10% to up to 25% of the electricity required of a regular domestic household dwelling. Turbines for residential scale use are usually between 7 feet (2 m) to 25 feet (8 m) in diameter and produce electricity at a rate of 900 watts to 10,000 watts at their tested wind speed. Building integrated wind turbine performance can be enhanced with the addition of an aerofoil wing on top of a roof mounted turbine. Solar water heaters, also called solar domestic hot water systems, can

[3] Turby wind turbine: can because horizontal axis (HAWT) types cannot change their pitch to face the wind directly. The turbine measures 2.0m (6'7") in diameter by 2.9m (9'6") high (including generator), and weighs 136 kg (300 lb). It is specified to generate power in winds of between 4 m/s (9 mph, 7.8kts) and 14 m/s (31 mph, 27.2kts), and can survive winds of 55 m/s (123 mph, 107kts). The rated power at 14 m/s is 2.5 kW (3.35 hp). The AC output from the synchronous generator is rectified to DC, then inverted to AC at 230V 50 Hz. Core International developed the turbine

Quoted summary: One source states the [1 amount of power produced by a wind turbine is proportional to the cube of the wind speed] . Other sources state [2 Turbines for residential scale use] [2 produce electricity at a rate of 900 watts to 10,000 watts] , and [3 is specified to generate power in winds of between 4 m/s (9 mph, 7.8kts) and 14 m/s (31 mph, 27.2kts)] .

Question: a component is what?

[1] Modular programming: in Dart, Go or Java) is sometimes used instead of module. In other implementations, this is a distinct concept; in Python a package is a collection of modules, while in Java 9 the introduction of the new module concept (a collection of packages with enhanced access control) is planned. Furthermore, the term "package" has other uses in software (for example .NET NuGet packages). A component is a similar concept, but typically refers to a higher level; a component is a piece of a whole system, while a module is a piece of an individual program. The scale of the term

[2] Physical body: the system at a point in time changes from identifying the object to not identifying it. Also an object's identity is created at the first point in time that the simplest model of the system consistent with perception identifies it. An object may be composed of components. A component is an object completely within the boundary of a containing object. In classical mechanics a physical body is collection of matter having properties including mass, velocity, momentum and energy. The matter exists in a volume of three-dimensional space. This space is its extension. Under Newtonian gravity the gravitational field further away

Quoted summary: [1 A component is a similar concept, but typically refers to a higher level; a component is a piece of a whole system, while a module is a piece of an individual program] in terms of [1 Modular programming] . Whereas in the [2 Physical body] , a [2 component is an object completely within the boundary of a containing object] .

Question: how can energy efficiency be improved?

[1] HVAC. forced air systems, which are now widely used in churches, schools and high-end residences, are A drawback is the installation cost, which can be slightly higher than traditional HVAC systems: Energy efficiency can be improved even more in central heating systems by introducing zoned heating. This allows a more granular application of heat, similar to non-central heating systems. Zones are controlled by multiple thermostats. In water heating systems the thermostats control zone valves, and in forced air systems they control zone dampers inside the vents which selectively block the flow of air. In this case, the control system is very

[2] Electrical efficiency: is valid only for non-reactive source and load impedances. High efficiency is particularly relevant in systems that can operate from batteries. Inefficiency may require weighing the cost either of the wasted energy, or of the required power supply, against the cost of attaining greater efficiency. Efficiency can usually be improved by choosing different components or by redesigning the system. Inefficiency probably produces extra heat within the system, which must be removed if it is to remain within its operating temperature range. In a climate-controlled environment, like a home or office, heat generated by appliances may reduce heating costs or increase

[3] Efficient energy use: Efficient energy use, sometimes simply called energy efficiency, is the goal to reduce the amount of energy required to provide products and services. For example, insulating a home allows a building to use less heating and cooling energy to achieve and maintain a comfortable temperature. Installing LED lighting, fluorescent lighting, or natural skylight windows reduces the amount of energy required to attain the same level of illumination compared to using traditional incandescent light bulbs. Improvements in energy efficiency are generally achieved by adopting a more efficient technology or production process or by application of commonly accepted methods

[4] Nuclear power phase-out: non-renewable energy sources. Renewable energy encompasses wind, biomass (such as landfill gas and sewage gas), hydropower, solar power (thermal and photovoltaic), geothermal, and ocean power. These renewable sources serve as alternatives to conventional power generation from thermal power stations run on nuclear or fossil fuels. A significant part of energy transition is reducing consumption by energy conservation and improvements in energy efficiency, an example is improved insulation for buildings; or improved energy efficiency by cogeneration of heat and power. Smart meters are able to charge higher prices at the time consumption peaks during the day, thereby causing electricity demand to

Quoted summary:

Table D.2: Example 2-shot QuoteSum prompt with two examples from the training set retrieved by question similarity. The demonstration summaries for each example are randomly picked out of the available references.

Instruction: Write a high-quality answer for the given question using only the provided search results and cite them properly using [1][2][3].

Question: how much power does a wind turbine produce?

[1] Compact wind acceleration turbine: It is generally thought that since the amount of power produced by a wind turbine is proportional to the cube of the wind speed, any acceleration benefit is potentially statistically significant in the economics of wind. As noted though this is an inaccurate as it ignores the impact of the exit to area ratio and is therefore an apples to oranges comparison. In the case of a typical CWAT/DAWT the power result in perfect theoretical operation once adjusted for the area of the shroud is actually the square of the velocity at the rotor. As the CWAT/DAWT diverges from theoretical function the power increase drops significantly according

[2] Sustainable architecture: roof ledge. Small-scale rooftop wind turbines have been known to be able to generate power from 10% to up to 25% of the electricity required of a regular domestic household dwelling. Turbines for residential scale use are usually between 7 feet (2 m) to 25 feet (8 m) in diameter and produce electricity at a rate of 900 watts to 10,000 watts at their tested wind speed. Building integrated wind turbine performance can be enhanced with the addition of an aerofoil wing on top of a roof mounted turbine. Solar water heaters, also called solar domestic hot water systems, can

[3] Turby wind turbine: can because horizontal axis (HAWT) types cannot change their pitch to face the wind directly. The turbine measures 2.0m (6'7") in diameter by 2.9m (9'6") high (including generator), and weighs 136 kg (300 lb). It is specified to generate power in winds of between 4 m/s (9 mph, 7.8kts), and 14 m/s (31 mph, 27.2kts), and can survive winds of 55 m/s (123 mph, 107kts). The rated power at 14 m/s is 2.5 kW (3.35 hp). The AC output from the synchronous generator is rectified to DC, then inverted to AC at 230V 50 Hz. Core International developed the turbine

Answer: One source states the amount of power produced by a wind turbine is proportional to the cube of the wind speed [1]. Other sources state Turbines for residential scale use produce electricity at a rate of 900 watts to 10,000 watts , and is specified to generate power in winds of between 4 m/s (9 mph, 7.8kts) and 14 m/s (31 mph, 27.2kts) [2] [3].

Question: a component is what?

[1] Modular programming: in Dart, Go or Java) is sometimes used instead of module. In other implementations, this is a distinct concept; in Python a package is a collection of modules, while in Java 9 the introduction of the new module concept (a collection of packages with enhanced access control) is planned. Furthermore, the term "package" has other uses in software (for example .NET NuGet packages). A component is a similar concept, but typically refers to a higher level; a component is a piece of a whole system, while a module is a piece of an individual program. The scale of the term

[2] Physical body: the system at a point in time changes from identifying the object to not identifying it. Also an object's identity is created at the first point in time that the simplest model of the system consistent with perception identifies it. An object may be composed of components. A component is an object completely within the boundary of a containing object. In classical mechanics a physical body is collection of matter having properties including mass, velocity, momentum and energy. The matter exists in a volume of three-dimensional space. This space is its extension. Under Newtonian gravity the gravitational field further away

Answer: A component is a similar concept, but typically refers to a higher level; a component is a piece of a whole system, while a module is a piece of an individual program in terms of Modular programming [1]. Whereas in the Physical body , a component is an object completely within the boundary of a containing object [2].

Question: how can energy efficiency be improved?

[1] HVAC. forced air systems, which are now widely used in churches, schools and high-end residences, are A drawback is the installation cost, which can be slightly higher than traditional HVAC systems: Energy efficiency can be improved even more in central heating systems by introducing zoned heating. This allows a more granular application of heat, similar to non-central heating systems. Zones are controlled by multiple thermostats. In water heating systems the thermostats control zone valves, and in forced air systems they control zone dampers inside the vents which selectively block the flow of air. In this case, the control system is very

[2] Electrical efficiency: is valid only for non-reactive source and load impedances. High efficiency is particularly relevant in systems that can operate from batteries. Inefficiency may require weighing the cost either of the wasted energy, or of the required power supply, against the cost of attaining greater efficiency. Efficiency can usually be improved by choosing different components or by redesigning the system. Inefficiency probably produces extra heat within the system, which must be removed if it is to remain within its operating temperature range. In a climate-controlled environment, like a home or office, heat generated by appliances may reduce heating costs or increase

[3] Efficient energy use: Efficient energy use, sometimes simply called energy efficiency, is the goal to reduce the amount of energy required to provide products and services. For example, insulating a home allows a building to use less heating and cooling energy to achieve and maintain a comfortable temperature. Installing LED lighting, fluorescent lighting, or natural skylight windows reduces the amount of energy required to attain the same level of illumination compared to using traditional incandescent light bulbs. Improvements in energy efficiency are generally achieved by adopting a more efficient technology or production process or by application of commonly accepted methods

[4] Nuclear power phase-out: non-renewable energy sources. Renewable energy encompasses wind, biomass (such as landfill gas and sewage gas), hydropower, solar power (thermal and photovoltaic), geothermal, and ocean power. These renewable sources serve as alternatives to conventional power generation from thermal power stations run on nuclear or fossil fuels. A significant part of energy transition is reducing consumption by energy conservation and improvements in energy efficiency, an example is improved insulation for buildings; or improved energy efficiency by cogeneration of heat and power. Smart meters are able to charge higher prices at the time consumption peaks during the day, thereby causing electricity demand to

Answer:

Table D.3: An example 2-shot QSum-S prompt that converts QuoteSum examples to ALCE-style sentence-level citations, and also provides the ALCE instructions. This question is similar to the one in Table D.2 showing the matching Qsum prompt for reference.

Question	*where is st mark's church located?*
Sources	[1] Church of St: Luke and The Epiphany (Philadelphia). southwestern edge of development within the original boundaries of Philadelphia (it would take almost another 10 years before an Episcopal church was organized across Broad Street and below Chestnut Street – St. Mark's Church at 16th and Locust Streets). These men advanced the money to build the church and were repaid through the sale of pews. Several sites were considered including the corner Broad and Locust Streets and the corner of 13th and Spruce Streets. These locations were rejected due to cost. The committee chose the present site of the church and selected Philadelphia architect Thomas S. Stewart to design [2] Charles Eamer Kempe: or wheatsheaf, taken from Kempe's own coat of arms. The mid-Victorian period were important years in the history of the design of English churches and Kempe's influence is found in numerous examples, many in his home county of Sussex which has 116 examples of his work. The works at St Mark's, Staplefield near Horsham, West Sussex dating from 1869 are regarded as especially important, representing the earliest of three known examples of Kempe's wall painting. They contain key elements of Kempe's figurative work. The angels holding the scroll are magnificently apparelled and the borders of their cloaks are embellished with [3] St: Mark's Cathedral, Bangalore. St. Mark's Cathedral is named after Saint Mark (believed to be the first gospel writer) and is located at the west end of Mahatma Gandhi Road, MacIver Town, Bangalore . Its architecture is inspired by the 17th century St Paul's Cathedral. The church serves as the Cathedral (Ecclesia Matar) of the Church of South India, Karnataka Central Diocese. Found in 1808, the cathedral celebrated its 200 years bicentenary in 2007-8. The cathedral is open for all people irrespective of religion or faith. The large grounds of the cathedral with numbers a trees offers a peaceful refugee from [4] Saint Mark's Coptic Orthodox Church (London): Saint Mark's Coptic Orthodox Church is a church in Kensington, London , located at the south-west corner of the intersection of Allen Street and Scarsdale Villas. It is dedicated to the founder of the Church of Alexandria, Mark the Evangelist. The building was previously Saint John's Presbyterian Church and was opened in 1863. Purchased from the Scottish Presbyterian church in 1975, Saint Mark's was the first Coptic Orthodox church in Europe. It was the seed church of the Coptic Orthodox Church in Britain and Ireland, first served by three hieromonks who later became members of
PaLM2 Unicorn 5-shot SEMQA prompt	Sources state St. Mark's Church is at 16th and Locust Streets , St Mark's, Staplefield is near Horsham, West Sussex , St. Mark's Cathedral is at the west end of Mahatma Gandhi Road, MacIver Town, Bangalore , and Saint Mark's Coptic Orthodox Church is in Kensington, London .
PaLM2 Bison ALCE prompt	St. Mark's Church is located at the west end of Mahatma Gandhi Road, MacIver Town, Bangalore [3].

Table E.1: Example generated answers with a 5-shot PaLM2 Unicorn model using our SEMQA prompt with QuoteSum examples, and a 4-shot PaLM2 Bison using the ALCE prompt. Colored spans show the spans that the SEMQA model marked as explicit extractions from the respective source, by color. In this example, the ALCE output answered only based on source number 3. The SEMQA answer covers all 4 passages with in-line quotes to attribute the respective sources.