

# Synthetic Multimodal Question Generation

Ian Wu<sup>\*◇</sup> Sravan Jayanthi<sup>\*◇†</sup> Vijay Viswanathan<sup>+</sup> Simon Rosenberg<sup>◇</sup>

Sina Pakazad<sup>◇</sup> Tongshuang Wu<sup>+</sup> Graham Neubig<sup>+</sup>

<sup>◇</sup>C3 AI    <sup>+</sup>Connectly AI    <sup>+</sup>Carnegie Mellon University

{ian.wu, simon.rosenberg, sina.pakazad}@c3.ai, sravan@connectly.ai  
{sherryw, gneubig}@cs.cmu.edu, vijayv@andrew.cmu.edu

## Abstract

Multimodal Retrieval Augmented Generation (MMRAG) is a powerful approach to question-answering over multimodal documents. A key challenge with evaluating MMRAG is the paucity of high-quality datasets matching the question styles and modalities of interest. In light of this, we propose **SMMQG**, a synthetic data generation framework. SMMQG leverages interplay between a retriever, large language model (LLM) and large multimodal model (LMM) to generate question and answer pairs directly from multimodal documents, with the questions conforming to specified styles and modalities. We use SMMQG to generate an MMRAG dataset of 1024 questions over Wikipedia documents and evaluate state-of-the-art models using it, revealing insights into model performance that are attainable only through style- and modality-specific evaluation data. Next, we measure the quality of data produced by SMMQG via a human study. We find that the quality of SMMQG-generated synthetic data is on par with the quality of the crowdsourced benchmark MMQA and that downstream evaluation results using both datasets strongly concur.

## 1 Introduction

Following the increased adoption of Retrieval Augmented Generation (RAG) (Lewis et al., 2021) for text-based question-answering (QA), there has been much interest in extending RAG to the multimodal setting (Chen et al., 2022; Chang et al., 2022; Lin and Byrne, 2022; Yasunaga et al., 2023). In multimodal RAG (MMRAG), QA is performed by a large language model (LLM) or large multimodal model (LMM) grounded in sources that span modalities such as text, tables and images. As with RAG, MMRAG has the potential to increase answer quality and transparency when compared with

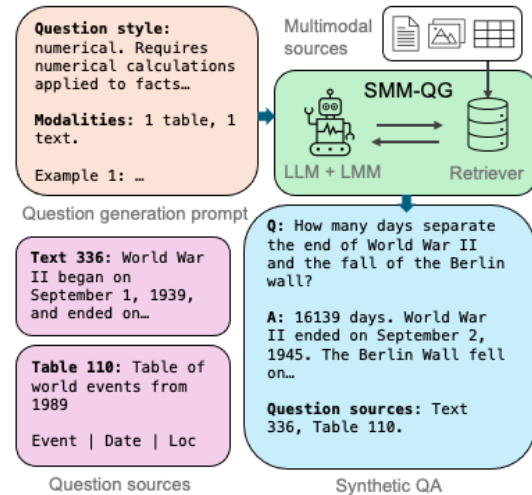


Figure 1: **An overview of SMMQG.** Given user-provided question style and modality requirements, SMMQG selects question sources and produces questions and answers. The questions are grounded in the selected question sources, and adhere to the question and modality requirements.

closed-book QA, where models directly answer questions without access to external knowledge.

A major challenge encountered when implementing MMRAG systems is evaluation, which is typically done using fixed benchmark datasets. These datasets consist of (*source(s)*, *question*, *answer*) tuples that enable separate evaluation of the retriever and the QA model. Some representative datasets include MMQA (Talmor et al., 2021), MMMU (Yue et al., 2023) and WebQA (Chang et al., 2022).

The problem with this approach is that we cannot tailor the evaluation questions to our desired specifications. We identify two aspects of the evaluation questions we may wish to control. The first aspect is the *question style*. The style of a question determines the reasoning abilities required to answer it, and certain models perform better on certain question styles than others. Examples of question styles include mathematical (Hendrycks

<sup>†</sup>Work done while at C3 AI.

et al., 2021; Yu et al., 2023), multi-hop (Yang et al., 2018) and extractive (Rajpurkar et al., 2016) question styles. The modality of a question refers to the modality or modalities (e.g. text, table, image) of the sources required to answer it, and both retrieval (Wei et al., 2023) and QA (Jia et al., 2021; Chen et al., 2020) performance depends on the modalities of their inputs.

Given the sensitivity of MMRAG performance to question styles and modalities, we want our evaluation questions to match the styles and modalities of the questions our system is likely to encounter. Such a benchmark may not exist, making it difficult to perform the comprehensive evaluations needed to reveal important model deficiencies.

To address this issue, we introduce a method for **Synthetic Multimodal Question Generation** we call **SMMQG**. SMMQG is a synthetic data generation framework that leverages interplay between a retriever, an LLM and an LMM with in-context learning (Brown et al., 2020) to generate multimodal questions and answers based directly on input documents. Crucially, SMMQG enables fine-grained control over the styles and modalities of questions, and is capable of producing both unimodal and cross-modal questions. In order to demonstrate the utility of SMMQG, we first use it to create a dataset of 1024 questions and answers of various styles and modalities from Wikipedia documents. We then use this dataset to unearth novel style- and modality-specific insights into the MMRAG performance of various models. Such insights would be difficult to obtain without style- and modality-specific evaluation data, which SMMQG can produce in an automatic and scalable fashion.

One concern with synthetic data generation is that the resulting data is of low quality. To assuage this concern and verify that SMMQG produces high quality data, we conduct a human study to measure the quality of our synthetic dataset. We find that our dataset’s quality is on par with or better than that of popular crowdsourced benchmark MMQA (Talmor et al., 2021) when measured along five different metrics. We also show that downstream evaluation results obtained using our SMMQG dataset display strong concurrence (Liu et al., 2023b) with those obtained using MMQA, demonstrating that our synthetic dataset can be used in place of MMQA for model selection.

## 2 Problem Setting

### 2.1 Multimodal Sources

We define the multimodal sources  $S$  to include the text passages, tables, and images extracted from one or more specified documents. Multimodal sources are parsed from a document as a pre-processing step.<sup>1</sup> Text sources consist of chunked text passages and table sources consist of pipe-separated strings that are prepended with titles. Image sources consist of an image, its caption and an image *verbalisation*. The image verbalisation is a text description of the image generated using an image-captioning model or LMM. Verbalisations are useful because they allow text-based models to search and reason over images, as seen in Liu et al. (2023c). See Appendix A for examples of sources and Appendix B for details of our image verbalisation strategy.

### 2.2 Formulation

SMMQG takes three inputs. These are (1) multimodal sources  $S$ , which serve to provide the context for question generation (2) question style  $v$ , which is a description of the question style along with examples, and (3) modality requirements  $M$ . This is a 3-tuple of integers  $M = (m_{\text{text}}, m_{\text{table}}, m_{\text{image}})$ .  $M$  is used to indicate the modalities of the generated questions:  $M = (2, 1, 0)$ , for example, indicates that our generated question should be a cross-modal text-table question with two text and one table sources.

SMMQG jointly produces three outputs. These are (1) the synthetic question  $q$ , with style dependent on  $v$  (2) a long-form answer  $a$  to the question and (3) references to the question sources  $Z$ , where  $z_i \in S$ .  $q$  is only answerable using information derived from *every* source in  $Z$ . The modalities of the question sources must match  $M$ : if  $M = (1, 1, 0)$ , for example, then  $|Z| = 2$  and  $z_1$  and  $z_2$  must be text and table modalities in some order.

## 3 Method

### 3.1 SMMQG

SMMQG is composed of five steps, as illustrated in Figure 2. The output of the first three steps is a set of *candidate sources*  $\tilde{Z}$ , where  $\tilde{z}_i \in S$ . These consist of semantically-related sources connected

<sup>1</sup>We are agnostic to the parsing strategy (Zhao et al., 2023), so long as contents of different modalities are parsed as separate sources.

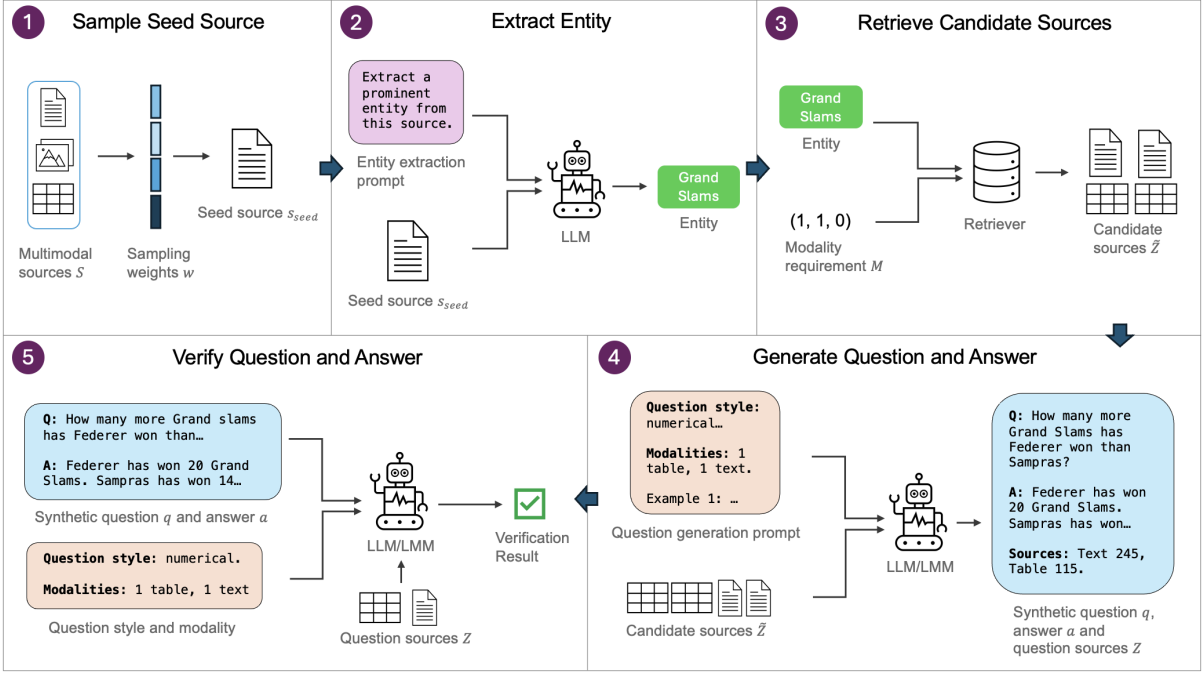


Figure 2: **SMMQG consists of five steps.** (1) A seed source is sampled from the sources  $S$ . (2) An entity is extracted from the seed source. (3) Candidate sources are retrieved from  $S$  using the entity from step 2 as the query. (4) The question generation model chooses the question sources from amongst the candidate sources, and uses these to produce the question and answer. (5) The model is asked to verify that the generated question adheres to the desired question style and modalities, and that the generated answer correctly answers the question.

by an entity. The fourth step is responsible for generating  $q$  and  $a$  and for choosing  $Z \subseteq \tilde{Z}$ . Thematic unity is maintained by the fact that all candidate sources are semantically related, which enables the generation of meaningful multi-source and potentially cross-modal questions. The fifth step is responsible for validating  $q$  and  $a$ .

**Step 1: Sample Seed Source** The goal of this step is to locate a seed source  $s_{seed} \in S$ . The most straightforward way to select  $s_{seed}$  is to choose one by uniformly sampling over  $S$ . However, we find that many sources are outliers that are unrelated to other sources and do not reflect the main topics found in the documents. Building coherent multi-source questions from such sources is difficult, as there are often no related candidate sources to choose from.

To correct for this, we introduce weights  $w_i$ . The probability of sampling  $s_i$  as  $s_{seed}$  is

$$p_{s_i} = \frac{\exp(-\beta w_i)}{\sum_j \exp(-\beta w_j)}$$

where  $\beta$  is a temperature parameter, which we set to 0.1 based on manual inspection of resulting outputs. The weights  $w_i$  are the average cosine-distance of the  $k_{seed}$ -nearest neighbours of  $s_i$  in embedding

space. Given some embedding model  $E$ ,

$$w_i = \frac{1}{k_{seed}} \sum_{s_j \in k_{seed} \text{nn}(s_i)} \text{dist}(E(s_i), E(s_j))$$

We use the E5-Large embedding model (Wang et al., 2024) in our experiments, with  $k_{seed} = 5$ .

**Step 2: Extract Entity** In this step, we use GPT-4-Turbo (OpenAI, 2024) to extract a prominent entity (e.g. “tennis”, “Japan”, “machine learning”) from the seed source via three-shot prompting. For image sources, we provide the LLM with the image verbalisation and the image caption, which we find captures image entities sufficiently well. We also find that using a high temperature of 1.0 improves the diversity of extracted entities, which in turn improves the diversity of generated questions. See Appendix C for our entity extraction prompt.

**Step 3: Retrieve Candidate Sources** In this step, we retrieve candidate sources  $\tilde{Z}$  using the E5-Large retriever, with the extracted entity from Step 2 as the query. The candidate sources are therefore all semantically related through the entity.

We also define an integer  $k_{modality}$ . For every modality  $i \in \{\text{text, table, image}\}$ , we retrieve the top  $m_i k_{modality}$  candidate sources of that modal-

ity. For example, given  $M = (1, 2, 0)$  we retrieve  $k_{\text{modality}}$  text and  $2k_{\text{modality}}$  table sources respectively. We set  $k_{\text{modality}} = 2$ , as we find that providing a choice of candidate sources improves the quality of generated questions. For retrieving images, we rely on image verbalisations. We find that this text-only retrieval approach improves over the use of multimodal retrieval with CLIP (Radford et al., 2021) embeddings, likely due to the prevalence of long text and table sources.

**Step 4: Question Generation** We now pass  $\tilde{Z}$  to the question-generation model, which is an LLM or LMM, depending on whether there are image candidate sources present. We use GPT-4-Turbo (OpenAI, 2024) as both in our experiments. In addition to  $\tilde{Z}$ , the question-generation model is also given the task instruction, the question style and its description  $v$ , the modality requirements  $M$ , and three style-specific few-shot examples. The task instruction asks the model to choose question sources  $Z$  from  $\tilde{Z}$ , adhering to  $M$ , and to then generate  $q$  and  $a$  using  $Z$ , following  $v$ . The model is also asked to produce references to  $Z$  by outputting their source IDs.<sup>2</sup> We instruct the model to refuse requests if it does not believe that a coherent question adhering to  $v$  and  $M$  can be generated. See Appendix C for our question-generation prompts.

**Step 5: Question Verification** We perform three checks and reject samples that fail any of these. Firstly, we cross-reference the modalities of the chosen question sources with  $M$  and reject samples where the source modalities do not match the requirements. For example, if  $M = (0, 1, 1)$ , we check that the modalities of  $Z$  are exactly 1 table and 1 image. Secondly, we pass  $q$  to an LLM and prompt it to verify that the question adheres to the specified question style. Lastly, we pass  $q$ ,  $a$  and  $Z$  to an LLM or LMM and prompt it to verify that (a)  $a$  is correct given  $q$  and  $Z$  and (b) every source in  $Z$  is required to answer  $q$ . We perform both the second and third steps of QA verification with a single call to GPT-4-Turbo. See Appendix D for our QA verification prompts.

### 3.2 Multi-hop QA Generation

Although Step 4 of SMMQG can be used to generate multi-hop questions, we propose an alternative version that produces higher quality multi-hop questions more consistently. We start by generat-

<sup>2</sup>Candidate sources that are not chosen could be used as *distractors* (Talmor et al., 2021), as they are semantically similar to the chosen sources.

ing two *intermediate questions* and their respective answers and question sources: for the first intermediate question, we prompt the model to generate a *question about the entity* extracted in Step 2, given some subset of  $\tilde{Z}$ . For the second question, we prompt the model to generate a question *where the answer is the same entity*, given the remaining  $\tilde{Z}$ . When we generate cross-modal multi-hop questions,  $\tilde{Z}$  are split by modality. When we generate unimodal multi-hop questions,  $\tilde{Z}$  are split randomly.

Next, we prompt the model to *combine* the intermediate questions and answers to form a multi-hop question and answer. The multi-hop question sources are the union of the question sources chosen in the intermediate steps. See Appendix E for examples of the question combination prompts and a diagram illustrating multi-hop question generation.

## 4 Experiments

### 4.1 Building a Synthetic Multimodal Wikipedia QA Dataset

We use SMMQG to build a QA dataset over Wikipedia documents. We use this dataset to facilitate experiments (1) demonstrating the utility of SMMQG as a framework for generating style- and modality-specific datasets (Section 4) and (2) verifying that SMMQG produces high quality data (Section 5). We use the text passages, tables and images gathered by Talmor et al. (2021) for MMQA as our sources. These sources include approximately 57,000 captioned images, 232,000 text passages and 12,000 tables, and were scraped from the 2020-01-01 English Wikipedia dump.

We preprocess the dataset by removing text passages with lengths of less than 200 characters, as we find that short passages rarely contain enough information to generate good questions. We generate a total of 1024 QA samples across five different question styles covering all pairwise modality combinations.<sup>3</sup> These are summarised in Table 1, and further examples are detailed in Appendix A. We choose these five question styles because (1) they test for a diverse set of reasoning abilities (2) they are well-represented across the QA literature (Khashabi et al., 2018; Rajpurkar et al., 2016; Yang et al., 2018; Dua et al., 2019).

<sup>3</sup>SMMQG can be used to generate multimodal questions with three or more question sources (Luo et al., 2023). We leave exploration of this to future work.

| Style                  | Description  | Example  | Modalities   | #   |
|------------------------|--|--|--|-----|
| Information Extraction | Requires extracting and returning information from a single source.  | What was the target age group for the NBC Kids programming block?  | Text, Table, Image   | 158 |
| Compare Contrast       | Requires comparing and contrasting two closely related entities or topics.                                   | Compare and contrast the logos of the French Open and the US Open tennis championships.  | Text-Text, Text-Table, Text-Image, Table-Table, Table-Image, Image-Image | 227 |
| Numerical              | Requires numerical calculations applied to facts extracted from the sources.                                 | How many years after the 1st Academy Awards did Dustin Hoffman receive his first nomination?   | Text, Table, Text-Text, Text-Table, Table-Table                          | 208 |
| Compound               | Two loosely connected information extraction questions, separated by "and".                                  | What is the scale used for passer rating in the NFL, and who is the all-time passing yards leader in professional football league history? | Text-Text, Text-Table, Text-Image, Table-Table, Table-Image, Image-Image | 206 |
| Multi-hop              | Requires first resolving an implicit sub-question, the answer to which is used to resolve the full question. | What color is the background of the flag of the location where the Echigo-Tsumari Art Triennial is held?                                   | Text-Text, Text-Table, Text-Image, Table-Table, Table-Image              | 225 |

Table 1: **Summary of our SMMQG-generated Wikipedia QA dataset.** This dataset consists of 1024 QA pairs, spanning five question styles and all pairwise modality combinations. Examples are taken directly from our dataset.

We also generate an additional dataset from different source documents to demonstrate that SMMQG is compatible with more domain-specific sources. See Appendix J for details.

## 4.2 Retriever and QA Model Evaluation

We use our dataset to evaluate the performance of three retrievers<sup>4</sup> and eight LLM + LMM combinations. For retrieval, we evaluate BM25, E5-Large (Wang et al., 2024) (both text-based) and OpenCLIP (Cherti et al., 2022) – an open-source implementation of CLIP offering improved performance – using recall@5 and recall@10 as our metric. For the text-based retrievers, we rely on captions and verbalisations for retrieving over images.

We evaluate the LLM + LMM combinations by conditioning answer generation on the SMMQG questions and question sources<sup>5</sup> and then scoring the predictions against the answers using GPT-4-Turbo as a judge, which has been shown to produce judgements that correlate strongly with human judgements (Zheng et al., 2023; Kim et al., 2024). The judge is asked to score the predicted answer on a three-point scale given the SMMQG-generated answer and the question sources. For QA, we use the LMM when the question sources contain at least one image; otherwise we use the LLM. An example of the judge prompt is provided in Appendix F. In addition to GPT-4-Turbo-judge scores, we also report GPT-3.5-Turbo-judge

<sup>4</sup>We use ChromaDB (Chroma, 2022) as the vector database for storing and retrieving embeddings.

<sup>5</sup>This evaluation setup is easier than end-to-end evaluation where the retrieved sources are used for QA, but has the advantage of allowing for independent evaluation of the retrieval and QA components.

(Brown et al., 2020), BERTScore (Zhang et al., 2020) and ROUGE (Lin, 2004) scores. We include these results in Appendix G.

We assess the open-source model combinations Vicuna-7b-v1.5 + LLaVA-v1.5-7b, its 13b counterpart (Zheng et al., 2023; Liu et al., 2023a), and Qwen-Chat + Qwen-Chat-VL (Bai et al., 2023a,b). The LLaVA-v1.5 models are LMMs finetuned from Vicuna-v1.5 (Peng et al., 2023), and both Qwen models are 7B parameter models finetuned from Qwen-LM (Bai et al., 2023a). We also assess the proprietary multimodal models GPT-4-Turbo (OpenAI, 2024), Gemini Pro 1.0 (Google, 2024) and the Claude 3 family (Anthropic, 2024), which are trained to process text-only inputs in addition to multimodal inputs and so can act as both the LLM and LMM. See Appendix I for further details.

**Evaluation Results** Retriever evaluation results are shown in the top subtables of Tables 2 and 3. We find that the text-based E5 retriever outperforms BM25 and OpenCLIP across all question styles, and is especially strong on abstractive styles (numerical, compare contrast, multi-hop), although we stress that our evaluation process is likely biased towards E5 as it was used in the question generation process itself. We also find that OpenCLIP significantly outperforms text-based retrieval for pure image retrieval questions, but underperforms text-based retrieval for all other modalities. Our findings underscore the importance of using robust, general-purpose multimodal embedders that are capable of performing consistently across diverse modalities, and we hope that our work can be used to support the development of such models.

The QA evaluation results are shown in the bot-

|                          |        | Info Extraction | Compare Contrast | Numerical   | Compound    | Multi-hop   |
|--------------------------|--------|-----------------|------------------|-------------|-------------|-------------|
| BM25                     | top-5  | 53.2            | 37.7             | 56.8        | 37.4        | 35.8        |
|                          | top-10 | 56.6            | 44.5             | 63.6        | 42.0        | 39.3        |
| E5                       | top-5  | 65.8            | 65.9             | 76.9        | 52.4        | 44.4        |
|                          | top-10 | <b>67.8</b>     | <b>72.2</b>      | <b>82.6</b> | <b>57.5</b> | <b>50.9</b> |
| OpenCLIP                 | top-5  | 61.1            | 38.3             | 51.6        | 36.9        | 18.0        |
|                          | top-10 | 67.1            | 48.9             | 59.5        | 43.3        | 24.0        |
| Vicuna-7b + LLaVA-7b     |        | 81.6            | 35.9             | 13.3        | 52.9        | 49.3        |
| Vicuna-13b + LLaVA-13b   |        | <b>89.1</b>     | <b>55.5</b>      | 33.2        | <b>73.5</b> | 60.2        |
| Qwen-Chat + Qwen-VL-Chat |        | 87.5            | 48.0             | <b>42.7</b> | 67.7        | <b>65.6</b> |
| Gemini Pro 1.0           |        | <b>96.3</b>     | 51.4             | 64.0        | 89.8        | 74.3        |
| Claude 3 Haiku           |        | 90.4            | 54.6             | 44.8        | 75.7        | 54.9        |
| Claude 3 Sonnet          |        | 93.0            | 67.0             | 63.3        | 79.2        | 72.7        |
| Claude 3 Opus            |        | 96.1            | <b>81.3</b>      | <b>77.7</b> | <b>90.8</b> | <b>88.7</b> |
| GPT-4-Turbo              |        | 99.3            | 89.9             | 85.3        | 96.8        | 96.2        |

Table 2: **Retrieval and QA evaluation results by question style.** The *top* subtable contains retrieval recall@5 and recall@10. The *bottom* subtable contains GPT-4-Turbo-judge scores for QA, where the top section contains results for open-source models, the middle contains results for proprietary models, and the bottom contains results for GPT-4-Turbo. We denote in **bold** the best models in each category. For retrieval, E5 achieves the best performance across all styles. Amongst open-source QA models, no single model dominates, while for proprietary QA models excluding GPT-4-Turbo, Claude 3 Opus demonstrates the best performance.

tom subtables of Tables 2 and 3. To start, we see that GPT-4-Turbo outperforms all other models across all question styles and modalities. This strength may be attributable to (1) GPT-4-Turbo being genuinely strong (2) GPT-4-Turbo being the evaluation data generation model, which filters out many questions that the model is unable to answer (3) GPT-4-Turbo being itself used as the judge, as LLM judges favour their own outputs (Koo et al., 2023).

Amongst the other models, Claude 3 Opus generally performs best, although further analysis yields novel style- and modality-specific insights. Firstly, Gemini Pro 1.0 performs similarly to Claude 3 Opus on extractive question styles (information extraction, compound), but is weaker on abstractive styles, where it is comparable to the smaller Claude 3 models. Secondly, the gap between open-source and proprietary models is small for extractive question styles but large for abstractive styles, with open-source models being especially poor on numerical and compare contrast questions. Comparing open-source models, we find that Qwen-Chat + Qwen-VL-Chat performs better on numerical and multi-hop questions than Vicuna-13b + LLaVA-13b despite being significantly smaller.

On the modality front, we learn that Gemini Pro 1.0 performs strongly for questions containing image sources, and is stronger than even Claude 3 Opus for unimodal image questions, although it suffers from weaker text and table reasoning abilities. Claude 3 Haiku, meanwhile, is surprisingly

poor at table QA (comparable to open-source models), but makes up for this with superior image reasoning capabilities.

In summary, we demonstrate that SMMQG can generate question style- and modality-specific evaluation datasets. Such datasets reveal important insights into the style- and modality-specific strengths and weaknesses of retrievers and QA models that would otherwise remain hidden.

## 5 Assessing Data Quality

### 5.1 MMQA

In the following sections, we assess the quality of our SMMQG-generated dataset and use MMQA (Talmor et al., 2021) as a reference to better understand the significance of our results. MMQA is a crowdsourced benchmark constructed over the same documents as ours. It contains both uni- and cross-modal questions, and provides long-form answers.

We use MMQA as our reference dataset for three reasons. Firstly, the modalities present in MMQA overlap exactly with our modalities of interest, unlike datasets such as WebQA (Chang et al., 2022) and OK-VQA (Marino et al., 2019), which contain only text and images. Secondly, MMQA is built on source documents spanning a diverse set of domains (Wikipedia), unlike domain-specific datasets such as BioASQ (Krithara et al., 2023) (biomedical) and MMMU (Yue et al., 2023) (exams and textbooks). Using it as a reference is therefore

|                          |        | Txt          | Tab         | Im          | Txt-Txt     | Txt-Tab     | Txt-Im      | Tab-Tab     | Tab-Im      | Im-Im       |
|--------------------------|--------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| BM25                     | top-5  | 92.9         | 50.0        | 12.0        | 76.9        | 39.8        | 43.4        | 20.0        | 20.2        | 15.4        |
|                          | top-10 | 95.2         | 54.5        | 14.0        | 86.1        | 45.5        | 46.5        | 26.5        | 24.5        | 18.4        |
| E5                       | top-5  | 98.8         | 79.5        | 20.0        | 91.5        | 60.2        | 52.3        | 51.0        | 34.6        | 24.3        |
|                          | top-10 | <b>100.0</b> | <b>81.8</b> | 22.2        | <b>96.3</b> | <b>70.3</b> | <b>56.2</b> | <b>60.2</b> | <b>38.5</b> | 26.5        |
| OpenCLIP                 | top-5  | 78.6         | 30.7        | 84.0        | 55.8        | 28.0        | 36.4        | 21.1        | 19.2        | 39.7        |
|                          | top-10 | 85.7         | 36.4        | <b>90.2</b> | 63.6        | 36.2        | 46.1        | 28.6        | 23.6        | <b>49.3</b> |
| Vicuna-7b + LLaVA-7b     |        | 65.9         | 43.5        | <b>74.0</b> | 52.6        | 28.0        | 60.1        | 12.5        | 43.2        | 46.8        |
| Vicuna-13b + LLaVA-13b   |        | 77.6         | 60.2        | <b>74.0</b> | <b>70.9</b> | <b>63.8</b> | 61.9        | 50.4        | 43.7        | <b>47.1</b> |
| Qwen-Chat + Qwen-VL-Chat |        | <b>82.3</b>  | <b>66.7</b> | 65.0        | 67.0        | 56.2        | <b>69.6</b> | <b>51.3</b> | <b>50.5</b> | 30.9        |
| Gemini Pro 1.0           |        | 84.4         | 71.3        | <b>97.0</b> | 78.6        | 57.3        | 82.9        | 59.1        | <b>84.6</b> | 80.4        |
| Claude 3 Haiku           |        | 78.4         | 61.6        | 87.8        | 61.3        | 42.4        | 75.7        | 36.2        | 63.1        | 80.1        |
| Claude 3 Sonnet          |        | 92.7         | 75.7        | 86.7        | 86.5        | 65.8        | 76.3        | 51.8        | 72.1        | 80.4        |
| Claude 3 Opus            |        | <b>96.1</b>  | <b>91.2</b> | 92.0        | <b>90.9</b> | <b>81.9</b> | <b>84.9</b> | <b>75.9</b> | 83.7        | <b>86.0</b> |
| GPT-4-Turbo              |        | 97.4         | 92.7        | 98.0        | 96.5        | 94.1        | 93.4        | 86.2        | 93.3        | 89.0        |

Table 3: **Retrieval and QA evaluation results by modality.** The *top* subtable contains retrieval recall@5 and recall@10. The *bottom* subtable contains GPT-4-Turbo-judge scores for QA, where the top section contains results for open-source models, the middle contains results for proprietary models, and the bottom contains results for GPT-4-Turbo. We denote in **bold** the best models in each category. For retrieval, OpenCLIP performs best for pure image retrieval, while E5 performs best for other modalities. Amongst open-source QA models, no single model dominates, while for proprietary QA models excluding GPT-4-Turbo, Claude 3 Opus generally demonstrates the best performance.

likely to yield more generalisable results. Lastly, the *single modality* and *compose*<sup>6</sup> questions from MMQA are stylistically very similar to the info extraction and multi-hop questions from our SM-MQG dataset; the presence of overlapping question styles enables direct comparison between datasets.

## 5.2 Human Study

We conduct a human study to directly evaluate the quality of SMMQG-produced data. We randomly draw 300 samples from our dataset, along with 60 single modality and 60 compose samples from the train set of MMQA, ensuring even distribution over modalities. We combine the samples and ask crowdworkers to rate them along five metrics:

- **Question Fluency** (5-likert): How fluent is the question?
- **Question Style Faithfulness** (Yes/No): Is the question faithful to the question style?
- **Source Relevance** (Yes/No): Are all the sources relevant to answering the question?
- **Answerability** (Yes/No): Is the question answerable using only information in the sources?
- **Answer Correctness** (Yes/No): Is the answer correct given the sources?

<sup>6</sup>*Single modality* questions are the TextQ, TableQ and ImageQ and *compose* questions the Compose(TextQ, TableQ), Compose(TableQ, TextQ), Compose(ImageQ, TextQ) and Compose(ImageQ, TableQ) question types from MMQA.

We compute the average rating for each metric for each dataset and report the results in Table 4. See Appendix L for further details on our human study methodology and for style-specific results. We employ the Mann-Whitney U test<sup>7</sup> (for Question Fluency) and Fisher’s exact test<sup>8</sup> (for the remaining metrics) to determine statistical significance. Our findings are as follows:

**SMMQG achieves high question style faithfulness.** This finding suggests that SMMQG can reliably produce questions with styles based on user specifications.

**SMMQG questions are more fluent than MMQA questions across comparable styles.** Any results we obtain better reflect the true reasoning capabilities of the model, as there is less interference caused by poor phrasing. However, we do not assess the ability of the model to handle poorly-phrased questions, which may reflect real user queries (Kwiatkowski et al., 2019).

**SMMQG question sources are highly relevant.** One source of error in SMMQG arises when the question-generation model selects question sources that the generated question is not based on. Our human study results address this concern.

<sup>7</sup>This non-parametric test compares differences between two independent groups when the dependent variable is either ordinal or continuous but not normally distributed.

<sup>8</sup>This test is used to determine if there are nonrandom associations between two categorical variables. In our case, the variables are dataset (SMMQG/MMQA) and human judgement for a given metric (Yes/No).

|            | Q. Fluency   | Q. Style Faithfulness | Source Relevance | Answerability | A. Correctness |
|------------|--------------|-----------------------|------------------|---------------|----------------|
| SMMQG      | 4.53         | 98.3                  | 93.0             | 94.7          | 92.7           |
| MMQA       | 3.68         | 96.7                  | 85.8             | 85.8          | 80.0           |
| $\Delta$   | <b>+0.85</b> | +1.6                  | +7.2             | <b>+8.9</b>   | <b>+12.7</b>   |
| $p$ -value | <0.001*      | 0.77                  | 0.07             | 0.02*         | 0.001*         |

Table 4: **Human study results.** We denote statistically significant  $\Delta$  in **bold** and  $p$ -values with  $p \leq 0.05$  using \*. These results show that our SMMQG dataset quality is on par with (and sometimes exceeds) the quality of MMQA.

**SMMQG questions are highly likely answerable given the question sources.** Another potential source of error arises when the question-generation model generates questions that are not answerable given its own selected question sources. Our study shows that SMMQG questions are actually statistically significantly more likely answerable than MMQA questions.

**SMMQG answers are highly likely to be correct.** A high level of answer correctness reduces the noise associated with incorrect labels leading to incorrect assessments of the predicted answers, and our SMMQG answers are statistically significantly more likely to be correct.

### 5.3 Measuring Concurrence

|           | $\tau$ | $p$ -value |
|-----------|--------|------------|
| Retrieval | 0.87   | 0.02*      |
| QA        | 0.86   | 0.002*     |

Table 5: **Concurrence of SMMQG for retrieval and QA.** We report Kendall’s tau and its associated  $p$ -values on the ranked list of retrieval and QA model evaluation results and find strong concurrence ( $\tau > 0.8$ ). We use \* to denote statistical significance ( $p \leq 0.05$ ).

We compute the *concurrence* (Liu et al., 2023b) between our SMMQG dataset and MMQA. The motivation for this is as follows: if we assume that MMQA is a useful evaluation dataset, and if our SMMQG dataset discriminates between models in the same way as MMQA, then we can reasonably conclude that our SMMQG dataset is also a useful evaluation dataset (Viswanathan et al., 2023).

We randomly draw 150 information extraction and 150 multi-hop samples from our SMMQG dataset, evenly distributed over modalities. We also randomly draw 150 single modality and 150 compose samples from MMQA, again distributed evenly over modalities. We then run evaluation using the methodology described in Section 4.2 (see Appendix K for these results) and calculate concurrence by computing Kendall’s tau on the ranked lists of these results. We report our findings in Table 5.

We find that MMQA and SMMQG strongly concur ( $\tau > 0.8$ , Liu et al. (2023b)) for both retrieval and QA. This implies that our SMMQG dataset can be used in-place of MMQA to discriminate between models (at least for the two common question styles), further validating its quality.

## 6 Related Work

**Multimodal QA Benchmarks** Many existing benchmarks rely on human annotators to handcraft questions and answers over fixed sets of documents. MMQA (Talmor et al., 2021) curate question-answer pairs by combining compositional question templates with crowdsourcing. This limits the complexity and diversity of generated questions (Chen et al., 2022). MMMU (Yue et al., 2023), which was constructed from college-level education material, was costly to curate, requiring the input of over 50 individuals. Other human-crafted multimodal QA benchmarks include WebQA (Chang et al., 2022), BioASQ (Krithara et al., 2023), ScienceQA (Lu et al., 2022), InfoSeek (Chen et al., 2023) and OK-VQA (Marino et al., 2019)

**Synthetic QA Generation** There exists a large body of work leveraging synthetic data to train and evaluate text-only QA systems. Puri et al. (2020) generate extractive QA data and use this to train a BERT-based (Devlin et al., 2019) model for QA. Shakeri et al. (2020) generate QA pairs using BART (Lewis et al., 2019) and use this for domain adaptation. Pan et al. (2020) propose a multi-step process to generate multi-hop questions, while Es et al. (2023) create a synthetic evaluation dataset for text-based RAG evaluation. Synthetic question generation has also been used to evaluate the quality of text summaries (Durmus et al., 2020; Wang et al., 2020). Note that, unlike SMMQG, none of the works discussed above allow for fine-grained control of question styles.

**Synthetic Multimodal QA Generation** Synthetic data generation has been studied in the context of visual QA (VQA). MultiQG-TI (Wang and Baraniuk, 2023) utilize an image-to-text model and OCR to create text descriptions of images that are



combined with text passages and then passed to an LLM for QA generation. [Patel et al. \(2020\)](#) build a diverse synthetic QA dataset from images and their associated metadata using an image-to-text model. These datasets concern QA over images and possibly image-text pairs only, and do not address text-only or table modalities. Furthermore, they do not enable control over question styles.

## 7 Conclusion

We introduce SMMQG, a framework for generating synthetic multimodal questions and answers grounded in multimodal documents that adhere to user-specified question styles and modalities. We demonstrate that an SMMQG-generated evaluation dataset is able to reveal novel style- and modality-specific insights into the performance of state-of-the-art retrievers, LLMs and LMMs. Through a human study and by measuring dataset concurrence, we also show that the quality of data generated by SMMQG is on-par with the quality of data from crowdsourced benchmark MMQA and that our dataset can be used in place of MMQA for model selection. We hope that SMMQG will enable automatic, large-scale and tailored evaluation of MMRAG systems, thereby facilitating their adoption in practical applications.

## Limitations

In our work, we evaluate SMMQG on Wikipedia documents and use GPT-4-Turbo to generate questions belonging to five question styles. SMMQG performance may differ when these variables are altered. The impact of new question styles and documents on SMMQG performance depends largely on how well GPT-4-Turbo is able to reason over these new question styles and understand these documents. This highlights a limitation of our work: we presuppose the existence of a model capable of reasoning over our question style and understanding our document of choice. Nonetheless, when such a model does exist, SMMQG enables us to generate synthetic data that can be used to evaluate other, possibly weaker models. Another related limitation of our work is that we only assess the viability of E5-Large and GPT-4-Turbo as the SMMQG retriever and question-generation model, and it is possible that new challenges arise when other models are used. Another limitation is that we limit our study to use of SMMQG data for evaluation, even though it is possible to use it for training. We

encourage further research in this direction, but we cannot yet claim that SMMQG data is appropriate for model training. Finally, our work relies on the existence of a high-quality set of unpaired data sources (in our experiments, this consists of images, text passages, and tables). This may not be an appropriate assumption in all situations; we did not test the ability of our dataset generation method to generalize to noisy data sources.

## Potential Risks

LLMs and LMMs are known to generate false, harmful and biased material. As SMMQG leverages LLMs and LMMs, it may potentially generate false, harmful and biased questions and answers. We also note that we have only explored the use of SMMQG for assessing the QA performance of MMRAG, and not its alignment to human values and preferences. MMRAG systems may excel on SMMQG-generated datasets but nonetheless be misaligned.

## References

- Team Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. [Qwen technical report](#).
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

- Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. [Webqa: Multihop and multimodal qa](#).
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. [Murag: Multimodal retrieval-augmented generator for open question answering over images and text](#).
- Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. [HybridQA: A dataset of multi-hop question answering over tabular and textual data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.
- Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. [Can pre-trained vision and language models answer visual information-seeking questions?](#)
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. [Reproducible scaling laws for contrastive language-image learning](#).
- Chroma. 2022. Home | Chroma — docs.trychroma.com. <https://docs.trychroma.com/>. [Accessed 03-05-2024].
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#).
- Esin Durmus, 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*. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. [Ragas: Automated evaluation of retrieval augmented generation](#).
- Samantha Fowler, Rebecca Roush, and James Wise. 2013. *Concepts of Biology*. OpenStax, Houston, Texas. Textbook content produced by OpenStax is licensed under a Creative Commons Attribution License 4.0.
- Team Google. 2024. [Gemini: A family of highly capable multimodal models](#).
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the math dataset](#).
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and vision-language representation learning with noisy text supervision](#).
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#).
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. [Benchmarking cognitive biases in large language models as evaluators](#). *ArXiv*, abs/2309.17012.
- Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. [Bioasqqa: A manually curated corpus for biomedical question answering](#). *Scientific Data*, 10(1):170.
- 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.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- 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. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#).
- 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.
- Weizhe Lin and Bill Byrne. 2022. [Retrieval augmented visual question answering with outside knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#).

- Nelson F. Liu, Tony Lee, Robin Jia, and Percy Liang. 2023b. [Do question answering modeling improvements hold across benchmarks?](#)
- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023c. [Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval.](#)
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering.](#)
- Haohao Luo, Ying Shen, and Yang Deng. 2023. [Unifying text, tables, and images for multimodal question answering.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore. Association for Computational Linguistics.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. [Ok-vqa: A visual question answering benchmark requiring external knowledge.](#)
- Team OpenAI. 2024. [Gpt-4 technical report.](#)
- Liangming Pan, Wenhua Chen, Wenhua Xiong, Min-Yen Kan, and William Yang Wang. 2020. [Unsupervised multi-hop question answering by question generation.](#) *arXiv preprint arXiv:2010.12623*.
- Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, and Jason Williams. 2020. [Generating natural questions from images for multimodal assistants.](#)
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. [Instruction tuning with gpt-4.](#)
- Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. [Training question answering models from synthetic data.](#)
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision.](#)
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100,000+ questions for machine comprehension of text.](#)
- Siamak Shakeri, Cicero Nogueira dos Santos, Henry Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. [End-to-end synthetic data generation for domain adaptation of question answering systems.](#)
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. [Multimodalqa: Complex question answering over text, tables and images.](#)
- Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu, and Graham Neubig. 2023. [Prompt2model: Generating deployable models from natural language instructions.](#)
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. [Asking and answering questions to evaluate the factual consistency of summaries.](#)
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. [Text embeddings by weakly-supervised contrastive pre-training.](#)
- Zichao Wang and Richard Baraniuk. 2023. [Multiqgti: Towards question generation from multi-modal sources.](#) *arXiv preprint arXiv:2307.04643*.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhua Chen. 2023. [Uniir: Training and benchmarking universal multimodal information retrievers.](#)
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering.](#)
- Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. [Retrieval-augmented multimodal language modeling.](#)
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. [Metamath: Bootstrap your own mathematical questions for large language models.](#)
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhua Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.](#)
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert.](#)
- Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. [Retrieving multimodal information for augmented generation: A survey.](#)

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#)

## A SMMQG Dataset Examples

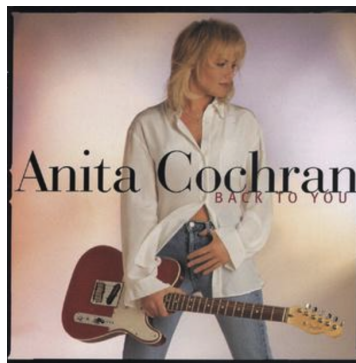
In this section, we provide examples of SMMQG-generated questions. Examples are cherry-picked to showcase questions of various question styles and modalities.

### A.1 Info Extraction, Image

**Question:** What musical instrument is Anita Cochran holding on the cover of her album "Back to You"?

**Answer:** Anita Cochran is holding a guitar on the cover of her album "Back to You".

**Question Source 1:**



**Image Caption:** Back to You (album)

**Image Verbalisation:** Anita Cochran, a country music singer, poses with her guitar for her album cover. The image features her holding the guitar and wearing a white shirt, showcasing her talent and passion for music.

### A.2 Numerical, Table

**Question:** What is the average position Pavlo Cherepin finished in during the 2014 rally season?

**Answer:** To calculate the average position, we add the positions from the 2014 season: 1 (Monte Carlo Rally) + 1 (Rally Mexico) + 3 (Rally d'Italia) + 1 (Rally Australia) and then divide by the number of events, which is 4. The calculation is  $(1+1+3+1)/4 = 6/4 = 1.5$ . Therefore, the average position Pavlo Cherepin finished in during the 2014 rally season is 1.5.

**Question Source 1:**

| # | Rally | Season             | Class | Pos.  | Co-driver | Car                   |                          |
|---|-------|--------------------|-------|-------|-----------|-----------------------|--------------------------|
| 0 | 1     | Rally Australia    | 2011  | PWRC  | 3         | Oleksandr Saliuk, Jr. | Mitsubishi Lancer Evo IX |
| 1 | 2     | Monte Carlo Rally  | 2014  | WRC 2 | 1         | Yuriy Protasov        | Ford Fiesta R5           |
| 2 | 3     | Rally Mexico       | 2014  | WRC 2 | 1         | Yuriy Protasov        | Ford Fiesta R5           |
| 3 | 4     | Rally d'Italia     | 2014  | WRC 2 | 3         | Yuriy Protasov        | Ford Fiesta RRC          |
| 4 | 5     | Rally Australia    | 2014  | WRC 2 | 1         | Yuriy Protasov        | Ford Fiesta RRC          |
| 5 | 6     | Rally d'Italia     | 2015  | WRC 2 | 1         | Yuriy Protasov        | Ford Fiesta RRC          |
| 6 | 7     | Rally Finland      | 2015  | WRC 2 | 3         | Yuriy Protasov        | Ford Fiesta RRC          |
| 7 | 8     | Rallye Deutschland | 2015  | RC 2  | 3         | Yuriy Protasov        | Ford Fiesta R5           |
| 8 | 9     | Rally Australia    | 2015  | WRC 2 | 2         | Yuriy Protasov        | Ford Fiesta RRC          |

**Table Title:** Pavlo Cherepin

### A.3 Multi-hop, Text-Image

**Question:** What colors are used in the logo of the NBA team that is based in Washington, D.C. and competes in the Eastern Conference Southeast Division?

**Answer:** The NBA team that is based in Washington, D.C. and competes in the Eastern Conference Southeast Division is the Washington Wizards. The Washington Wizards logo uses red, white, blue, and silver colors. Therefore, the colors used in the logo of the NBA team based in Washington, D.C. and competing in the Eastern Conference Southeast Division are red, white, blue, and silver.

**Question Source 1:** Washington Wizards: The Washington Wizards are an American professional basketball team based in Washington, D.C. The Wizards compete in the National Basketball Association (NBA), as a member of the league's Eastern Conference Southeast Division. The team plays its home games at the Capital One Arena, in the Chinatown neighborhood of Washington, D.C.

**Question Source 2:**



**Image Caption:** Washington Wizards

**Image Verbalisation:** Washington Wizards Basketball Team Logo

### A.4 Compare Contrast, Text-Text

**Question:** Compare and contrast the sculptures "The Spirit of Polonia" and "On Watch" in Milwaukee.

**Answer:** Both "The Spirit of Polonia" and "On Watch" are public sculptures located in Milwaukee, but they differ in their themes and representations. "The Spirit of Polonia" is a stainless steel sculpture with three rings symbolizing harmony, unity, and infinity, and a brass sphere representing earth, emphasizing abstract concepts. In contrast, "On Watch" is a life-size bronze sculpture depicting a policeman, a firefighter, and a child, focusing on the human figures and their roles as guardians of public safety. Additionally, "The Spirit of Polonia" is situated in front of the Milwaukee County Courthouse and includes a concrete pool, while "On Watch" is administered by the City of Milwaukee and features inscriptions and a plaque with dedications.

**Question Source 1:** On Watch: "On Watch" is a life-size bronze sculpture consisting of three figures. A Caucasian policeman and an African-American firefighter stand back-to-back holding a child wrapped in a blanket. The inscription on the southwest corner of the sculpture reads: WANNER 90. The inscription on the northwest corner reads: WACO FDR. There is a bronze plaque on the concrete base which reads: On Watch/ Dedicated to the guardians of public safety in the city of Milwaukee John Norquist, mayor Milwaukee Art Commission, David M. Wanner Sculptor 1990. The work is administered by the City of Milwaukee.

**Question Source 2:** "The Spirit of Polonia": This nine foot, five inch stainless steel sculpture has three rings meaning harmony, unity and infinity. While the brass sphere represents earth. Each ring is a different size having one inside the other, therefore having each one get smaller, then the "globe" is the smallest. These sculpture is surrounded by a sixteen-foot, five inch concrete pool. Both are in front of the Milwaukee County Courthouse.

## A.5 Compound, Text-Table

**Question:** What is the home stadium of the Miami Dolphins, and what was the attendance when the Baltimore Colts played against the Miami Dolphins in 1975?

**Answer:** The home stadium of the Miami Dolphins is Hard Rock Stadium, and the attendance when the Baltimore Colts played against the Miami Dolphins in 1975 was 61,986 on Week 10 and 59,398 on Week 13.

**Question Source 1: Hard Rock Stadium:** Hard Rock Stadium is a multipurpose football stadium located in Miami Gardens, Florida, a city north of Miami. It is the home stadium of the Miami Dolphins of the National Football League (NFL). Hard Rock Stadium also plays host to the Miami Hurricanes football team during their regular season. The facility also hosts the Orange Bowl, an annual college football bowl game. It was the home to the Florida Marlins of Major League Baseball (MLB) from 1993 to 2011.

**Question Source 2:**

| Week | Date           | Opponent                | Result  | Record | Game Site                     | Attendance |
|------|----------------|-------------------------|---------|--------|-------------------------------|------------|
| 0    | 1 September 21 | at Chicago Bears        | W 35–7  | 1–0    | Soldier Field                 | 54,152     |
| 1    | 2 September 28 | Oakland Raiders         | L 20–31 | 1–1    | Memorial Stadium              | 40,657     |
| 2    | 3 October 5    | at Los Angeles Rams     | L 13–24 | 1–2    | Los Angeles Memorial Coliseum | 62,491     |
| 3    | 4 October 12   | Buffalo Bills           | L 31–38 | 1–3    | Memorial Stadium              | 43,907     |
| 4    | 5 October 19   | at New England Patriots | L 10–21 | 1–4    | Schaefer Stadium              | 51,417     |
| 5    | 6 October 26   | at New York Jets        | W 45–28 | 2–4    | Shea Stadium                  | 55,137     |
| 6    | 7 November 2   | Cleveland Browns        | W 21–7  | 3–4    | Memorial Stadium              | 35,235     |
| 7    | 8 November 9   | at Buffalo Bills        | W 42–35 | 4–4    | Rich Stadium                  | 77,320     |
| 8    | 9 November 16  | New York Jets           | W 52–19 | 5–4    | Memorial Stadium              | 52,097     |
| 9    | 10 November 23 | at Miami Dolphins       | W 33–17 | 6–4    | Orange Bowl                   | 61,986     |
| 10   | 11 November 30 | Kansas City Chiefs      | W 28–14 | 7–4    | Memorial Stadium              | 42,122     |
| 11   | 12 December 7  | at New York Giants      | W 21–0  | 8–4    | Shea Stadium                  | 49,863     |
| 12   | 13 December 14 | Miami Dolphins          | W 10–7  | 9–4    | Memorial Stadium              | 59,398     |
| 13   | 14 December 21 | New England Patriots    | W 34–21 | 10–4   | Memorial Stadium              | 48,678     |

**Table Title:** 1975 Baltimore Colts season

## B Image Verbalisation

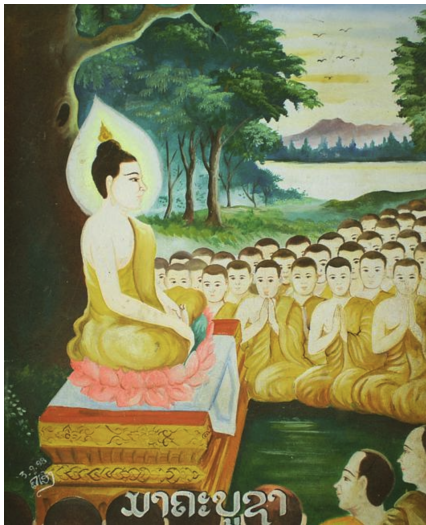
All image verbalisations used for our experiments were created with the llama-2-13b-chat-lightning-preview version of LLaVA, which was the most up-to-date LLaVA model at the time the verbalisations were created.

### B.1 Image Verbalisation Prompt

We pass the image along and any associated captions to LLaVA and use the following prompt to generate the image verbalisation:

Here is the image caption: {caption}.  
Here is an image. Describe the contents of the image, taking into account all portions. Try and be descriptive.

### B.2 Image Verbalisation Examples



**Image Caption:** Magha Puja

**Image Verbalisation:** In this painting, a large group of monks is gathered around a seated Buddha, listening intently to his teachings. The scene takes place in a serene outdoor setting, with a tree in the background. The image captures the essence of spirituality and the importance of learning from one's spiritual leader.



**Image Caption:** Fallout 4: Nuka-World

**Image Verbalisation:** A woman in a space suit is enjoying a thrilling ride on a roller coaster, with a bottle of soda in her hand. This image captures the excitement and enjoyment of the theme park experience.



## C SMMQG Prompts

### C.1 Entity Extraction Prompt

Here we provide the entity extraction prompts from Step 2 of SMMQG. We choose {num\_entities} to be 1 in our experiments, although this can be increased to further increase question diversity.

Identify up to {num\_entities} key themes or entities present in the passage. Make sure any entities you return are general and widely-known.

<few shot examples>

Passage: {passage}

Entities and terms:

### C.2 Question Generation Prompts

Here we provide the question generation prompts for Step 4 of SMMQG. For inputs containing text or tables only, the prompt is:

You are given a question style description, which describes the characteristics of a specific style of reading comprehension question, and some examples of questions of this style. You are also provided with modality requirements. Your task is to generate a question following the style specified in the question style description based on the input passages. Also generate an answer to the question and citations of the passages the answer is based on.

RULES

1. The question you generate MUST be based on one or more of the provided source passages.
2. The modality requirements constrain the passages you can choose as the source passages. For example, if the modality requirement is 1 text, 1 table, your question MUST be based on 1 text and 1 table passage exactly.
3. The question should not be answerable if any chosen source passage is removed.
4. If no modality requirements are given, you may generate questions based on any number of passages of any modality.
5. Do not mention the passage number in the question. Also do not explicitly mention the table, image or text.
6. Generate a natural sounding question that a reasonable human being might ask.
7. Produce a response in the following format: <question> | <answer> | <citation>. The citation should be a reference to all the source passages chosen.
8. Closely follow the template question description. If you cannot do this, say None.
9. If you cannot abide by any of the rules above, say None. It is preferable for you to say None than to risk breaking the rules.

{style\_prompt}

<few shot examples>

Passages: {enumerated\_passages}

Question | Answer | Citation:

For inputs containing images, we use the following prompt as the system prompt:

You are given a question style description, which describes the characteristics of a specific style of reading comprehension question. You are also provided with one or more captioned images, and possibly some additional text or table passages. In addition, you are also provided with modality requirements. Your task is to generate a question following the style specified in the question style description based on the input images, image captions and text or table passages. Also generate an answer to the question and citations of the images or passages the answer is based on.

#### RULES

1. The question you generate MUST be based on the provided images, image captions and passages (if provided). The subset of provided images and passages that the question is based on are the chosen source materials.
2. The modality requirements constrain the images and passages you can choose as the source materials. For example, if the modality requirement is 1 image, 1 text, your question MUST be based on 1 image (and its caption) and 1 text passage exactly.
3. The question should not be answerable if any of the chosen source materials are removed.
4. Generate a natural sounding question that a reasonable human being might ask.
5. Produce a response in the following format: <question> | <answer> | <citation>. The citations should be a reference to the images or passages chosen. You should images and passages by their number only
6. You MUST use the image captions of the source images you have chosen explicitly in the question. For example, if the caption says "Roger Federer", then "Roger Federer" must be explicitly stated in the question somewhere.
7. Closely follow the question style description. If you cannot do this, say None.

{style\_prompt}

Few shot examples, text and table passages and images are passed to GPT-4-Turbo as conversation turns via the chat completion API. Images are directly captioned, and non-image data is formatted according to the following template:

Passages: {enumerated\_passages}

Question | Answer | Citation:

### C.3 Question Style Prompts

The question style prompts contain descriptions  $v$  and are inserted into the question generation prompts at {style\_prompt}. Creating questions with new styles only involves writing new question style prompts along with corresponding few-shot examples. In this section, we provide the question style prompts used to generate our SMMQG Wikipedia dataset.

#### C.3.1 Info Extraction

Question style: Information extraction question. Simple question that can be answered by extracting a fact from a single passage or image, if provided. Examples of such questions: Who is the founder of Microsoft? On what date did the Battle of Stalingrad begin? Which two actors won the Academy Awards for Best Actor and Best Supporting Actor in 2012?

### **C.3.2 Compare Contrast**

Question style: Compare and contrast question. Requires making comparisons based on information from one or more passages or images, if provided. The two subjects being contrasted must belong to the same category and be directly comparable - being about the same topic is not enough. Do not create questions about subjects that are not closely related and do not belong to the same category - if you cannot be sure, say None. Examples of such question: Compare and contrast the embryonic development process of humans and monkeys. Compare and contrast the careers of tennis players Roger Federer and Rafael Nadal. Format your answer to this question like this: explain what the relationship between the two subjects are and why they are similar. Next, identify several comparable traits and explain how the traits of the subjects are similar or different. Avoid simply summarising one subject after the other - make sure to interweave both subjects in your answer.

### **C.3.3 Numerical**

Question style: Maths question. Requires calculation based on numbers from passages to determine the answer. Calculation must be used to answer the question, simple extraction of numbers is not enough. If there are no numbers mentioned in the passages, say None. If there are no calculations possible, say None. Examples of such questions include: how many more trophies did Barcelona win between 2000 and 2010 then Atleti? What was the percentage change in the number of Covid cases in Italy between March 2020 and September 2020?

### **C.3.4 Compound**

Question style: Compound question. Question is composed of two thematically related subquestions connected by "and". Examples of such question include: How many Grand Slams did Roger Federer win, and how many of those were at the US Open? Who is the first King of England, and what is the significance of the Crown Jewels to English royalty? What is a hurricane, and what weather event causes the most deaths in a typical year?

## D QA Verification

For inputs containing text or tables only, the prompt is:

You are given a question, an answer to that question, and some supporting passages. You are also given a question style, some information about it and some examples of questions in that style. Your job is to assess whether the question and answer passes or fails, based on 2 criteria:

Criterion 1. The answer to the question can be inferred from the supporting passages provided.

Criterion 2. The question closely matches the style of question specified by the question style.

If all 2 criteria are met, return Pass. If one or more conditions are not met, return Fail, along with the list of conditions that were not met.

<few shot examples>

Question: {question}

Answer: {answer}

Question Style: {style\_prompt}

Passages: {enumerated\_passages}

Assessment:

For inputs containing images, the following prompt is used as the system prompt:

You are given a question, an answer to that question, and some supporting images and (optionally) passages. You are also given a question style, some information about it and some examples of questions in that style. Your job is to assess whether the question and answer passes or fails, based on 2 criteria:

Criterion 1. The answer to the question can be inferred from the supporting images and (optionally) passages provided.

Criterion 2. The question closely matches the style of question specified by the question style.

If all 2 criteria are met, return Pass. If one or more conditions are not met, return Fail, along with the list of conditions that were not met.

Few shot examples, questions, answers, question styles, text and table passages and images are passed to GPT-4-Turbo as conversation turns via the chat completion API. Images are directly captioned, and non-image data is formatted according to the following template:

Question: {question}

Answer: {answer}

Question Style: {style\_prompt}

Passages: {enumerated\_passages}

Assessment:

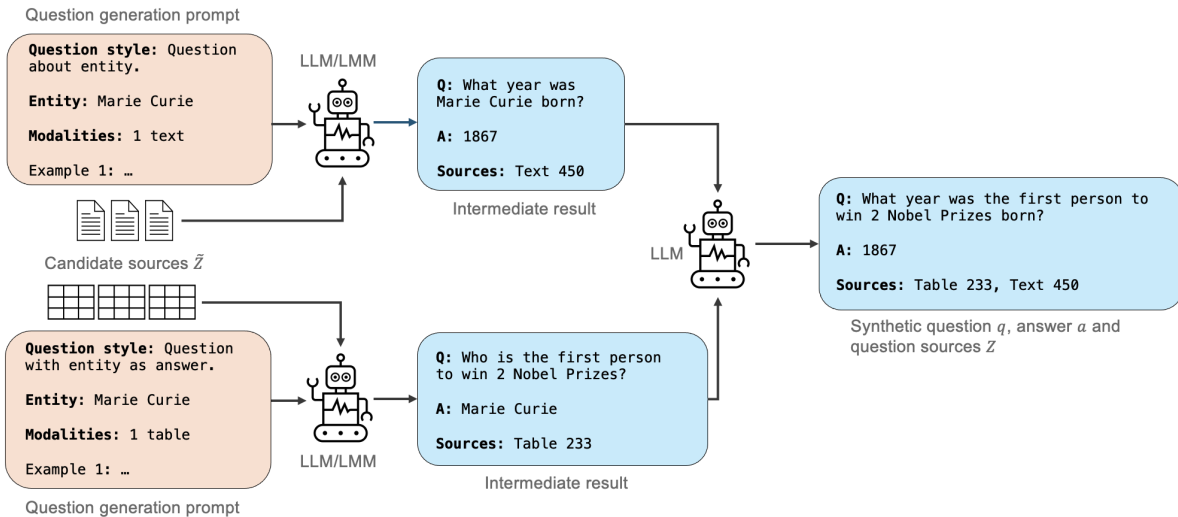


Figure 3: **Multi-hop Question Generation.** Instead of directly producing a question as in standard SMMQG, we first generate two intermediate questions. The first question is about the extracted entity, and the second has the extracted entity as the answer. These are then combined by an LLM or LMM to produce the final, multi-hop question.

## E Multi-hop Question Generation

### E.1 Intermediate Question Generation Prompts

The intermediate question generation steps for multi-hop question generation use the same question generation prompts as for standard question generation, but with specific question style prompts. Unlike before, we also insert the entity found in Step 2 into the question style prompt:

#### E.1.1 Question About Entity Prompt

Question style: Information extraction question about an entity: {entity}. Simple question that can be answered by extracting a fact from a single passage or image about the provided entity. The question you generate MUST contain a direct and explicit reference to the entity. Avoid generating questions asking what/who is entity e.g. do not ask "Who is Roger Federer", but ask something about Roger Federer explicitly. Also avoid generating questions not based on hard facts e.g. "what is company X known for", "what is person Y famous for". If you cannot generate a question about this entity, say None. Examples of such entity and question combinations: Entity: Microsoft. Question: Who is the founder of Microsoft? Entity: Battle of Stalingrad. Question: On what date did the Battle of Stalingrad begin? Entity: Academy Awards 2012. Question: Which actor won the Academy Award for Best Actor in 2012?

#### E.1.2 Question With Entity As Answer Prompt

Question style: Information extraction question where the answer is the provided entity: {entity}. Simple fact extraction question where the answer is the entity provided. The answer you give must be the entity itself, nothing more. If you cannot generate a question with the entity as the answer, say None. Try your best to create a question where the entity is likely the unique answer - this will require you to think about whether there might be other entities that can be used to answer the question. If such entities exist, ask a different question. Examples of such entity and question combinations: Entity: Microsoft. Question: Which company developed the Windows OS? Answer: Microsoft. Entity: Stalingrad. Question: What was the name of the city of Volgograd during the Second World War? Answer: Stalingrad.

## **E.2 Combination Prompt**

The combination prompt is used to combine the two intermediate questions into a multi-hop question via an LLM. For inputs containing text or tables only:

You are given an entity and two questions, answers and passages. The first question is a question about the entity, and the second question is a question with an answer that is the entity. Your task is to combine the questions into a single, multi-hop question, and combine the answers in order to create an answer to the multi-hop question. A multi-hop question is one that requires answering an implicit sub-question before the full question can be answered. For example, the question "Who is the wife of the 40th president of the US" is a multi-hop question, because it requires first answering "Who is the 40th president of the US?" and using that answer to answer the full question. You can always construct a multi-hop questions from two questions if one of the questions contains a direct reference to an entity that is also the answer to the second question. The procedure simply involves replacing the reference to the common entity in the first question with the second question.

#### RULES

1. Combine the questions into a multi-hop question, if possible.
2. Combine the answers into a single, answer that answers the multi-hop question step-by-step.
3. If the two provided questions do not fulfill the criteria for constructing a multi-hop question (one question contains an explicit reference to the entity, the other has as its answer the entity), return None.
4. Try and phrase the multi-hop question as naturally as possible without altering the validity of the answer, and make sure that the generated multi-hop question makes sense. If this is not possible and the question is deemed awkward, return None.
5. Return your multi-hop question and answer in the format: <multi-hop question | answer>.
6. If the second question can be answered by the first passage, or the first question can be answered by the second passage, return None.
7. If there no need to resolve the implicit question in order to resolve the full question, return None.
8. Check that the questions can be combined to form a meaningful question. Sometimes the individual questions are valid but cannot be combined. If the combined question is not valid, say None.
9. If you cannot fulfill any of the rules above, return None. It is preferable to return None than to break any of the rules.

<few shot examples>

Passages: {enumerated\_passages}

Question 1: {question\_1}

Question 2: {question\_2}

Answer 1: {answer\_1}

Answer 2: {answer\_2}

Entity: {entity}

Multi-hop Question | Answer:

For inputs containing images, the following prompt is used as the system prompt:

You are given an entity and two questions, two answers, some images and possibly some passages. The first question is a question about the entity, and should contain an explicit reference to the entity in it. The second question is a question with an answer that is the entity. Your task is to combine the questions into a single, multi-hop question, and combine the answers in order to create an answer to the multi-hop question. A multi-hop question is one that requires answering an implicit sub-question before the full question can be answered. For example, the question "Who is the wife of the 40th president of the US" is a multi-hop question, because it requires first answering "Who is the 40th president of the US?" and using that answer to answer the full question. You can always construct a multi-hop questions from two questions if one of the questions contains a direct reference to an entity that is also the answer to the second question. The procedure simply involves replacing the reference to the common entity in the first question with the second question.

#### RULES

1. Combine the questions into a multi-hop question, if possible.
2. Combine the answers into a single, answer that answers the multi-hop question step-by-step.
3. If the two provided questions do not fulfill the criteria for constructing a multi-hop question (one question contains an explicit reference to the entity, the other has as its answer the entity), return None.
4. If the answer to the first question (the one that should be containing an explicit reference to the entity) also has an answer that is the entity, return None.
5. If the entity provided is not the entity that bridges the two questions (i.e. is referenced in one and is the answer to the other), return None.
6. Try and phrase the multi-hop question as naturally as possible without altering the validity of the answer, and make sure that the generated multi-hop question makes sense. If this is not possible and the question is deemed awkward, return None.
7. Return your multi-hop question and answer in the format: <multi-hop question | answer>.
8. The passages/images associated with the first question are labelled/captions Passage 1/Image 1. The passages/images associated with the second question are labelled/captions Passage 2/Image 2. If Passage 1/Image 1 can be used to answer the second question, or vice versa, return None.
9. If you cannot fulfill any of the rules above, return None. It is preferable to return None than to break any of the rules.

Few shot examples, questions, answers, passages, entities and images are passed to GPT-4-Turbo as conversation turns via the chat completion API. Images are directly captioned, and non-image data is formatted according to the following template:



Passages: {enumerated\_passages}

Question 1: {question\_1}

Question 2: {question\_2}

Answer 1: {answer\_1}

Answer 2: {answer\_2}

Entity: {entity}

Multi-hop Question | Answer:

## F GPT-4-Turbo Judge

For inputs containing text or tables only:

You are a fair and unbiased judge. You have been given a question, a model answer to that question and some source passages. The source passages should contain all the information required to answer the question. You are then provided with a candidate answer to the question. Your objective is to score the candidate answer to the question. Use the sources and the model answer to better understand what the truth is and assign the candidate answer a score. Return an explanation followed by the score. Separate the explanation from the score with "Score:"

IMPORTANT: the model answer should only act as a guide. It is possible for the new answer to be different from the model answer but still be correct. You must think about the question and look at the source passages closely.

- Provide a score of 0 if the candidate answer is incorrect.
- Provide a score of 1 if the candidate answer is somewhat correct, but is missing something important or contains some minor inaccuracies.
- Provide a score of 2 if the candidate answer is correct.

<few shot examples>

Question: {question}

Candidate Answer: {candidate\_answer}

Model Answer: {model\_answer}

Passages: {passages}

Explanation and Score:

For inputs containing images, the following prompt is used as the system prompt:

You are a fair and unbiased judge. You have been given a question, a model answer to that question and one or more images and possibly some passages. The images and passages together form the sources, and this should contain all the information required to answer the question. You are then provided with a candidate answer to the question. Your objective is to score the candidate answer. Use the sources and the model answer to better understand what the truth is and assign the candidate answer a score. Return an explanation followed by the score. Separate the explanation from the score with "Score:".

IMPORTANT: the model answer should only act as a guide. It is possible for the candidate answer to be different from the model answer but still be correct. You must think about the question and look at the images and passages closely.

- Provide a score of 0 if the candidate answer is incorrect.
- Provide a score of 1 if the candidate answer is somewhat correct, but is missing something important or contains some minor inaccuracies.
- Provide a score of 2 if the candidate answer is correct.

Few shot examples, questions, passages, answers and images are passed to GPT-4-Turbo as conversation turns via the chat completion API. Images are directly captioned, and non-image data is formatted according to the following template:

```
Question: {question}
Candidate Answer: {candidate_answer}
Model Answer: {model_answer}
Passages: {passages}

Explanation and Score:
```

## **G Evaluation Results with Additional Metrics**

In addition to GPT-4 scores, we also report evaluation results computed using GPT-3.5-Turbo-judge, BERTScore and ROUGE scores. For GPT-3.5-Turbo, we ask the judge to compare the model answer (reference) against the predicted answer given the question and to score the answer as correct, incorrect or partially correct. Our GPT-3.5-Turbo judge prompt is as follows:

You are a fair judge. You are provided with a question, a reference answer and a prediction. Decide whether the prediction is correct based on the reference answer on a scale of 0 to 2. If the prediction is incorrect, return a 0. If the prediction is partially correct, return a 1. If the prediction is correct, return a 2. Do not use your own knowledge. Use only the reference.

Example 1

Question: Who was the F1 champion in 2019?

Reference: Lewis Hamilton

Prediction: Seb Vettel

Score: 0

Example 2

Question: Explain the purpose of dropout in neural networks.

Reference: Dropout is used for regularisation. It helps prevent overfitting.

Prediction: Reduction in overfitting via regularisation.

Score: 2

Example 3

Question: Compare and contrast the careers of Federer and Nadal.

Reference: Federer has won 20 grand slams, and Nadal has won 22. Federer retired in 2021, whereas Nadal currently still plays tennis.

Prediction: Federer won 20 grand slams and Nadal won 22.

Score: 1

Input

Question: {question}

Reference: {reference}

Prediction: {prediction}

Score:

As GPT-3.5-Turbo lacks multimodal capabilities, we do not provide sources for decision-making - the judge assesses model performance based only on the model answer. For BERTScore, we use microsoft/deberta-xlarge-mnli as the BERTScore model and treat the model answer as the reference and the predicted answer as the candidate. For ROUGE evaluation, we compute ROUGE-1 (unigram overlap), again between the model answer and the predicted answer.

We report aggregate results for each model in Table 6 and report Kendall’s tau on the evaluation ranked lists against GPT-4-Turbo judge in Table 7

|                          | GPT-4-Turbo Judge | GPT-3.5-Turbo Judge | BERTScore   | ROUGE       |
|--------------------------|-------------------|---------------------|-------------|-------------|
| Vicuna-7b + LLaVA-7b     | 45.3              | 56.9                | 61.7        | 53.0        |
| Vicuna-13b + LLaVA-13b   | 61.3              | 68.6                | 69.0        | 61.3        |
| Qwen-Chat + Qwen-VL-Chat | 61.1              | 69.2                | 66.8        | 66.2        |
| Gemini Pro 1.0           | 73.9              | 77.1                | <b>70.2</b> | 63.3        |
| Claude 3 Haiku           | 62.7              | 71.5                | 62.9        | 71.9        |
| Claude 3 Sonnet          | 74.1              | 80.3                | 66.2        | 70.9        |
| Claude 3 Opus            | 86.5              | 82.7                | 68.2        | 72.0        |
| GPT-4-Turbo              | <b>93.4</b>       | <b>87.8</b>         | 69.8        | <b>77.9</b> |

Table 6: Evaluation results with other evaluation metrics. We report results aggregated over all question styles and question modalities.

|            | <b>GPT-3.5-Turbo Judge</b> | <b>BERTScore</b> | <b>ROUGE</b> |
|------------|----------------------------|------------------|--------------|
| $\tau$     | 0.93                       | 0.50             | 0.72         |
| $p$ -value | <0.001*                    | 0.11             | 0.014*       |

Table 7: Kendall’s tau and its associated  $p$ -value on the ranked lists of evaluation results against GPT-4-Turbo judge. We find that GPT-3.5-Turbo judge and ROUGE correlate well with GPT-4-Turbo judge.

## H Analysis of Question Style Overlap

In order to better understand how distinct our pre-defined question styles are, we analyse the level of question style overlap in our SMMQG-generated dataset. We use GPT-4-Turbo as a question style binary classifier, pass each QA pair along with descriptions of the question styles to it and ask whether the question belongs to a given style. We define questions with overlap as those that elicit a positive response from the classifier for two or more styles, and report the percentage of questions in our dataset that contain overlap in Table 8, breaking down results by the ground-truth style:

|                  | Info Extraction | Compare Contrast | Numerical | Compound | Multi-hop |
|------------------|-----------------|------------------|-----------|----------|-----------|
| <b>% Overlap</b> | 4.3             | 0.0              | 4.9       | 2.8      | 20.0      |

Table 8: Percentage of questions with overlap by style. We define questions with overlap as those that are classified by GPT-4-Turbo as belonging to more than one question style. The only question style with significant overlap is the *Multi-hop* question style.

We see that overlap is insignificant for all but the multi-hop questions. Upon closer inspection, we notice that the multi-hop style commonly overlaps with the numerical and info extraction styles. Our explanation is that multi-hop reasoning inherently requires use of other reasoning skills because these other skills are needed to solve the individual subparts of multi-hop questions.

## I Model Details and Licences

| <b>Name</b> | <b>Hugging Face ID</b>                | <b>Licence</b> |
|-------------|---------------------------------------|----------------|
| E5-Large    | intfloat/e5-large-v2                  | mit            |
| OpenCLIP    | laion/CLIP-ViT-H-14-laion2B-s32B-b79K | mit            |

Table 9: Details for retriever models.

All model evaluation experiments were done using **greedy decoding**. All inference experiments with open-source models were done using 1 A100 GPU. We document model details and licenses in Tables 9, 10 and 11.

## J Additional Dataset

In addition to our main Wikipedia-derived SMMQG dataset, we also generate a separate QA dataset on a college-level biology textbook (Fowler et al., 2013) containing text and image sources.<sup>9</sup> We use the exact same SMMQG parameters and prompts as for the original dataset to generate this dataset. Our additional dataset contains 324 questions derived from 862 individual sources.

We conducted an additional human study to assess the quality of this dataset and thereby demonstrate that SMMQG generalises to other, more domain-specific sources than Wikipedia. We sampled 150 questions from our new dataset and asked crowdworkers to assess their quality, repeating the process described in Section 5.2. We report our findings in Table 12. We find that the quality of our new dataset remains high despite the use of a more domain specific dataset. This demonstrates that SMMQG generalises well beyond Wikipedia.

| Name                      | Details   | Licence           |
|---------------------------|---|-------------------|
| LLaVA-13b (Verbalisation) | liuhaotian/llava-llama-2-13b-chat-lightning-preview | LLAMA 2 Community |
| LLaVA-v1.5-7b             | liuhaotian/llava-v1.5-7b                            | LLAMA 2 Community |
| LLaVA-v1.5-13b            | liuhaotian/llava-v1.5-13b                           | LLAMA 2 Community |
| Vicuna-7b-v1.5            | lmsys/vicuna-7b-v1.5                                | LLAMA 2 Community |
| Vicuna-13b-v1.5           | lmsys/vicuna-13b-v1.5                               | LLAMA 2 Community |
| Qwen-Chat                 | Qwen/Qwen-7B-Chat                                   | Tongyi Qianwen    |
| Qwen-Chat-VL              | Qwen/Qwen-VL-Chat                                   | Tongyi Qianwen    |

Table 10: Details for open-source LLMs and LMMs.

| Name            | Details                                 |
|-----------------|---|
| GPT-4-Turbo     | gpt-4-1106-vision-preview               |
| Gemini Pro 1.0  | gemini-1.0-pro-vision-001               |
| Claude 3 Haiku  | anthropic.claude-3-haiku-20240307-v1:0  |
| Claude 3 Sonnet | anthropic.claude-3-sonnet-20240229-v1:0 |
| Claude 3 Opus   | anthropic.claude-3-opus-20240229-v1:0   |

Table 11: Details for API-based models.

|       | Q. Fluency | Q. Style Faithfulness | Source Relevance | Answerability | A. Correctness |
|-------|------------|-----------------------|------------------|---------------|----------------|
| SMMQG | 4.47       | 94.7                  | 86.4             | 91.2          | 91.2           |

Table 12: **Human study results on our additional dataset.** These results show that our biology SMMQG dataset quality is high as assessed by crowdworkers.

## K Concurrence Study

### K.1 Evaluation Results and Discussion

Tables 13 and 14 contain the SMMQG and MMQA evaluation results used in our concurrence study in Section 5.3. We see that information extraction and multi-hop SMMQG questions are in general easier to answer than their MMQA counterparts. One hypothesis is that this results from MMQA questions being less fluent and less answerable, as shown in Section 5.2. To investigate this, we remove all samples with unanswerable questions and with question fluencies of less than 4 from the datasets used in Section 5.3, and replace them with new samples that we manually validate. We re-run evaluation on this new dataset, but find that MMQA questions remain more challenging than SMMQG questions.

We conclude that, on comparable styles, our SMMQG questions are generally easier to answer than their MMQA counterparts. We hypothesise that this is the result of the question generation model picking the most “obvious” questions to ask, whereas crowdworkers may attempt to produce more novel questions. Nonetheless, we note that question style and modality have far more influence on the difficulty of questions, as evidenced by our results in Section 4.2, and that our SMMQG dataset holds similar discriminative power to MMQA despite being easier, as evidenced by our concurrence experiment results. Finally, we suggest that question difficulty may be increased via prompting through question style prompt  $v$ ; we leave exploration of this to future work.

## L Human Study

### L.1 Additional Results

We present additional human study results breaking down performance by question style in Table 15. We find that SMMQG-question quality is maintained across question styles.

<sup>9</sup><https://dept.clcillinois.edu/biodv/PrinciplesOfBiology.pdf>

|          |        | MMQA        | SMMQG       |
|----------|--------|-------------|-------------|
| BM25     | top-5  | 36.3        | <b>42.8</b> |
|          | top-10 | 42.1        | <b>45.0</b> |
| E5       | top-5  | <b>54.3</b> | 53.5        |
|          | top-10 | <b>60.1</b> | 56.7        |
| OpenCLIP | top-5  | 38.3        | <b>40.0</b> |
|          | top-10 | 43.7        | <b>45.3</b> |

Table 13: **Retrieval evaluation results** for our concurrence experiments in Section 5.3.

|                          | MMQA        | SMMQG       |
|--------------------------|-------------|-------------|
| Vicuna-7b + LLaVA-7b     | 55.0        | <b>68.0</b> |
| Vicuna-13b + LLaVA-13b   | 61.0        | <b>75.0</b> |
| Qwen-Chat + Qwen-VL-Chat | 65.7        | <b>78.5</b> |
| Gemini Pro 1.0           | 80.7        | <b>89.6</b> |
| Claude 3 Haiku           | 73.9        | <b>75.9</b> |
| Claude 3 Sonnet          | <b>87.3</b> | 86.9        |
| Claude 3 Opus            | <b>93.0</b> | 92.9        |
| GPT-4-Turbo              | 96.4        | <b>97.4</b> |

Table 14: **QA evaluation results** for our concurrence experiments in Section 5.3.

|                  | Q. Fluency   | Q. Style Faithfulness | Source Relevance | Answerability | A. Correctness |
|------------------|--------------|-----------------------|------------------|---------------|----------------|
| Info Extraction  | 4.73         | 100.0                 | 100.0            | 96.7          | 95.0           |
| Single Modality  | 3.77         | 100                   | 86.7             | 86.7          | 80.0           |
| $\Delta$         | <b>+0.96</b> | +0.0                  | <b>+13.3</b>     | +10.0         | <b>+15.0</b>   |
| $p$ -value       | <0.001*      | 1.0                   | 0.006*           | 0.09          | 0.025*         |
| Multi-hop        | 4.47         | 98.3                  | 91.7             | 100.0         | 95.0           |
| Compose          | 3.57         | 93.3                  | 85.0             | 85.0          | 80.0           |
| $\Delta$         | <b>+0.90</b> | +5.0                  | +6.7             | <b>+15.0</b>  | <b>+15.0</b>   |
| $p$ -value       | <0.001*      | 0.44                  | 0.39             | 0.003*        | 0.025*         |
| Compare Contrast | 4.52         | 98.3                  | 88.3             | 88.3          | 88.3           |
| Numerical        | 4.50         | 100.0                 | 88.3             | 91.7          | 91.7           |
| Compound         | 4.43         | 95.0                  | 96.7             | 96.7          | 93.3           |

Table 15: **Human study results by question style**. In the top and middle subtables, we compare *single modality* and *compose* questions from MMQA with SMMQG info extraction and multi-hop questions. In the bottom subtable, we report results for the remaining SMMQG styles. We denote statistically significant differences in **bold** and  $p$ -values with  $p \leq 0.05$  using \*.

## L.2 Additional Details

We recruited 25 crowdworkers for our human study via crowdsourcing platform Prolific ([www.prolific.com](http://www.prolific.com)). Each crowdworker was given a total of 90 minutes to read through the instructions and evaluate 20 samples, for which they were paid 17 GBP (with the added possibility of a bonus). We screened crowdworkers and selected only those (1) located in the US, UK, Ireland, Australia, New Zealand and Canada with English as their primary language and (2) possessing at least a Bachelor’s degree.

We conducted our human study via Google Forms. Each crowdworker was sent a link to a Google Form containing 20 samples and 3 attention checks, along with a further link to the instructions. The instructions were provided via a Google Doc that contained an in-depth explanation of the task, the evaluation metrics, the bonus structure, as well as three fully-worked examples.

Of the 20 samples, 3 were always test samples of varying difficulties. These were included in order to assess the quality of responses from each crowdworker. Crowdworkers that correctly labelled all 3 test samples were awarded a bonus of 8 GBP. We manually reviewed the responses of crowdworkers who incorrectly labelled 2 or more test samples and made the appropriate corrections (at no point during this review process were we made aware of the source of any given question). Of the 25 crowdworkers, 8 scored 3/3 on the test questions, 13 scored 2/3 and 4 scored 1/3 or below.

The remaining 17 samples in each form were drawn without replacement from a pool of SMMQG and MMQA samples. This pool was in turn composed of 120 MMQA and 300 SMMQG samples, with 60 samples from each of the two MMQA and five SMMQG styles present. These samples were drawn randomly without replacement from the full datasets. The final crowdworker was shown only 12 rather than 17 samples in order to maintain an even distribution over styles.

**QUESTION (5/20)**  
What role did Dana Carvey play in the film "Wayne's World"?

---

**ANSWER**  
Dana Carvey played the role of Garth Algar in the film "Wayne's World".

---

**QUESTION STYLE**  
Question that requires extracting information from a single source. No complex reasoning is required.

---

**SOURCE: Dana Carvey**

|    | Year | Title  | Role                     | Notes       |
|----|------|--|--------------------------|-------------|
| 0  | 1981 | Halloween II                                   | Assistant Barry McNichol |             |
| 1  | 1984 | This Is Spinal Tap                             | Mime Waiter              |             |
| 2  | 1984 | Racing with the Moon                           | Baby Face                |             |
| 3  | 1986 | Tough Guys                                     | Richie Evans             |             |
| 4  | 1988 | Moving   | Brad Williams            |             |
| 5  | 1990 | Opportunity Knocks                             | Eddie Farrell            |             |
| 6  | 1992 | Wayne's World                                  | Garth Algar              |             |
| 7  | 1993 | Wayne's World 2                                | Garth Algar              |             |
| 8  | 1994 | Clean Slate                                    | Maurice L. Pogue         |             |
| 9  | 1994 | The Road to Wellville                          | George Kellogg           |             |
| 10 | 1994 | Trapped in Paradise                            | Alvin Firpo              |             |
| 11 | 1996 | The Shot                                       | Himself                  | Cameo       |
| 12 | 1996 | Fire on the Track: The Steve Prefontaine Story | Himself                  | Documentary |
| 13 | 2000 | Little Nicky                                   | Referee                  | Cameo       |
| 14 | 2002 | The Master of Disguise                         | Pistachio Disguisey      | Also writer |
| 15 | 2010 | Presidential Reunion                           | George H. W. Bush        | Short film  |
| 16 | 2011 | Jack and Jill                                  | Crazy Puppeteer          | Cameo       |
| 17 | 2015 | Hotel Transylvania 2                           | Dana the Camp Director   | Voice       |
| 18 | 2016 | The Secret Life of Pets                        | Pops                     | Voice       |
| 19 | 2017 | Sandy Wexler                                   | Himself                  |             |
| 20 | 2017 | Becoming Bond                                  | Johnny Carson            | Documentary |
| 21 | 2017 | Too Funny to Fail                              | Himself                  | Documentary |
| 22 | 2019 | The Secret Life of Pets 2                      | Pops                     | Voice       |

Figure 4: An example section from a Google Form shown to one of the crowdworkers. This section contains the question, answer, question style and source.

How fluent is the question? A fluent question can be clearly understood, is free of grammatical mistakes, and is concise. \*

1 2 3 4 5

Not fluent at all      Completely fluent

Is the question faithful to the question style? A question that is faithful to the question style is one that meets the description of the specified question style. \*

YES

NO

Are all the sources provided relevant to answering the question? \*

YES

NO

Is the question answerable using information provided by the sources? \*

YES

NO

Is the answer correct? An answer is correct if it correctly answers the question posed, using logical reasoning and information from the sources. \*

YES

NO

Figure 5: An example section from a Google Form shown to one of the crowdworkers. This section contains the response section.



Crowdworkers were given these instructions via a Google Doc:

In this study, we compare various methods for creating questions and answers for reading comprehension tasks. We have used these methods to create some questions and answers. We would like your help deciding how good these questions and answers actually are. You will be given some questions and answers. Each question and answer pair is connected to one or two sources of information, which may take the form of text, tables or images, and a question style. Your job is to assess the question and answer along 5 criteria. There are 20 samples in total. We recommend spending on average between 3-4 minutes per sample. Some samples should require less time than this, and others more. Please read the following instructions carefully and make sure you understand the task before proceeding. We have factored in 15 minutes for you to read these instructions. The entire task, including instructions-reading, should take no more than 90 minutes to complete. Your completion code will be provided at the end of the Google form.

There are 5 different criteria to assess a sample on.

<continued on next page>

**Question Fluency:** Here we assess how fluent the question is. A completely fluent question is one that is free of grammatical mistakes, is phrased well and is concise. Your task is to decide how fluent the question is on a scale of 1 (incomprehensible) to 5 (fluent). Use your best judgment, but a question that is free of grammatical mistakes and is phrased well should be rated a 5, and a question that is incomprehensible should be rated a 1.

**Question Style Faithfulness:** The questions were created based on a question style, which describes the kind of question that should be created. The question style and its description is provided along with the question and answer. Your task is to decide if the question adheres to the question style specified.

**Source Relevance:** The questions were created based on one or two sources, and should be answerable using information provided in these sources. The sources may be a mix of text, table or image sources. Your task is to decide if all the sources contain information that is relevant to answering the question. If any source contains only irrelevant or tangentially relevant information, you should answer NO. If all the sources are all relevant and provide useful information, you should answer YES. If a source contains information about the correct topic but nothing useful can be extracted from it, you should still answer NO. If there are two sources, both of which contain the same useful information, you should answer YES.

**Answerability:** The questions were created based on one or two sources, and should be answerable using information provided in these sources. Your task is to decide if the sources provide enough information to produce a reasonable answer to the question. Do not use your own knowledge of the topic to decide whether or not the sources provide enough information. You should only look at the sources and decide if there is adequate information present to at least superficially answer the question. Reasoning from information contained in the sources is allowed, so long as all the facts can be derived from the sources.

**Answer Correctness:** The answer to the question is created along with the question. Here we assess whether the answer to the question is correct given the sources. Your task is to decide whether or not the answer to the question is correct using information from the sources. The answer does not need to be perfect for it to be correct. You should mark it as correct so long as it provides a reasonable answer to the question. Do not use your own knowledge of the topic to assess the answer. You should put yourself in the shoes of a reasonably intelligent person who has no knowledge of the topic at hand, and use only information contained in the sources, even if it is outdated or wrong. Reasoning from information contained in the sources is allowed, so long as all the facts can be derived from the sources.

<Working Examples>