QASA: Advanced Question Answering on Scientific Articles

Yoonjoo Lee [*1] **Kyungjae Lee** [*2] **Sunghyun Park** [2] **Dasol Hwang** [2] **Jaehyeon Kim** [2] **Hong-in Lee** [3] **Moontae Lee** [2 4]

Abstract

Reasoning is the crux of intellectual thinking. While question answering (QA) tasks are prolific with various computational models and benchmark datasets, they mostly tackle factoid or shallow QA without asking deeper understanding. Dual process theory asserts that human reasoning consists of associative thinking to collect relevant pieces of knowledge and logical reasoning to consciously conclude grounding on evidential rationale. Based on our intensive think-aloud study that revealed the three types of questions: surface, testing, and deep questions, we first propose the QASA benchmark that consists of 1798 novel question answering pairs that require full-stack reasoning on scientific articles in AI and ML fields. Then we propose the QASA approach that tackles the full-stack reasoning with large language models via associative selection, evidential rationale-generation, and systematic composition. Our experimental results show that QASA's full-stack inference outperforms the state-of-the-art INSTRUCTGPT by a big margin. We also find that rationale-generation is critical for the performance gain, claiming how we should rethink advanced question answering. The dataset is available at https://github.com/lgresearch/QASA.

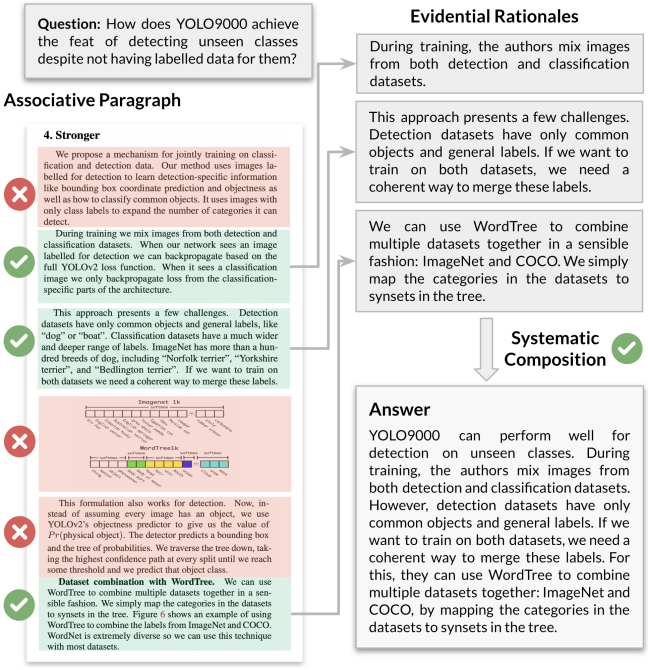

Figure 1. An example of QASA. A question that the reader/author asks about the paper while reading the paper. To formulate the answer, one classifies whether the paragraph contains evidence to answer the question. Evidential rationales are written for each evidential paragraph and are systematically composed into a comprehensive answer.

1. Introduction

Reasoning differentiates human intellectual capabilities from low-level intelligence. Dual process models theorize that cognitive reasoning is a two-stage process where the first stage performs associative thinking and the second stage performs logical reasoning (Wason & Evans, 1974; Tsujii & Watanabe, 2009; Evans, 2012). Within the context of Question Answering (QA), the first stage extracts associative pieces of knowledge based on lexical matching and other cognitive heuristics, inductively expanding potential evidences. Then the second stage consciously finds evidential rationales, deductively converging to the answer via systematic compositions of the evidences. This process uniquely characterizes advanced human reasoning, posing a non-trivial challenge to machine learning QA systems.

Reading Comprehension (RC) is one type of reasoning task that can formulate various questions and answers. SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), DROP (Dua et al., 2019), and Natural Questions (Kwiatkowski et al., 2019a) have been proposed. While com-

[*]Equal contribution [1]KAIST (Work done at LG AI Research) [2]LG AI Research [3]Yonsei University [4]University of Illinois Chicago. Correspondence to: Mootae Lee <moontae.lee@lgresearch.ai>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

peting on their model performance significantly improves machine answering capabilities, these datasets consist of factoid QAs mostly in the form of "what", "when", "where", or "who". Thus extracting short spans from the relevant context can easily provide correct answers, but the trained models can barely answer "how" and "why" questions.

Recent work on open-domain QA (Karpukhin et al., 2020; Guu et al., 2020; Liu et al., 2021; Izacard & Grave, 2021; Izacard et al., 2022) exploits the *Retrieve-then-read* approaches, where the system first retrieves relevant documents from a large corpus then reads out concrete answers. These approaches target shallow questions that are often inferable relying only on the first stage rather than jointly using the both stages. Some reasoning tasks like bAbI and its permuted version (Weston et al., 2015; Rajendran et al., 2018) require logically correct spatial reasoning. However, the artificial nature of their QAs minimally leverages the second stage as their reasoning tasks do require neither rich retrieval of associative information from the first stage nor systematic composition of the final answers (Lee et al., 2016).

Our think-aloud study reveals that reading scientific articles not only raises surface questions but also induces testing and deep questions that require full-stack reasoning. In addition, carefully answering surface questions turns out to involve both first and second stage reasoning, requiring significantly more elaborated efforts compared to what previous datasets and models implicitly assumed. To answer for such naturally advanced questions, we propose the **Q**uestion **A**nswering on **S**cientific **A**rticles (**QASA**), a novel QA benchmark and an approach that realize the full-stack cognitive reasoning from the first to the second stages. Our QASA benchmark differs from existing ones on the following aspects:

- Based on our think-aloud study, we design a schema for advanced questions as *surface*, *testing*, and *deep* questions, then collecting balanced QA pairs form the authors of research papers as well as from expert readers.
- We guide readers and authors to ask questions while reading the *whole paper* rather than gathering only extractive questions from paper abstracts.
- Readers and authors are asked to propose their *multi-faceted long-form* answers to the collected questions, then *composing* a comprehensive final passage than simply summarizing evidential rationales with added fluency.

Our QASA benchmark contains 1798 QA pairs on AI/ML papers where the questions are asked by regular readers of AI/ML papers and answered by AI/ML experts. Each paper has 15.1 questions on average, up to a maximum of 29 questions for a single paper. We collect 39.4% of deep reasoning level questions based on our own question schema. And, maximum 9 evidential rationales are leveraged to compose the final answer.

Method	Associative selection	Evidential rationale-generation	Systematic composition
QASPER	✓	✗	✗
ELI5	✗	✗	✗
ASQA	✓	✗	✓
AQuaMuSe	✗	✗	✓
QASA (ours)	✓	✓	✓

Table 1. Comparison of existing datasets and our QASA.

Our QASA approach models the full-stack reasoning process via state-of-the-art large language models. We decompose the process into three subtasks: *associative selection* (to extract relevant information from paragraphs), *evidential rationale-generation* (to grasp only evidential rationale from each extracted paragraph), and *systematic composition* (to stitch evidential rationales into a comprehensive answer without redundancy). Modeling each subtask by pretrained large language model with existing datasets, we demonstrate that our best test-bed outperforms the state-of-the-art InstructGPT (OpenAI's text-davinci-003) by 5.11 Rouge-1 points. We further verify that directly generating an answer from selected paragraphs causes performance drop, opening a crucial insight for tackling advanced question answering.

2. Related Work

The relevant research consists of three categories: QA for academic research papers, long-form QA, and query-based multi-document summarization. Table 1 highlights our method against existing approaches in each groups.

QA for Academic Research Papers Several datasets have been proposed for QA on academic research papers including emrQA (Pampari et al., 2018), BioRead (Pappas et al., 2018), and BioMRC (Stavropoulos et al., 2020). They automatically construct their QA examples by extracting entities and relations as well as structure knowledge resources. Thus these datasets would unlikely reflect real-world scenarios where users have more advanced and open-ended questions (Kwiatkowski et al., 2019b). Closest to our work, QASPER (Dasigi et al., 2021), consists of 5K QA on NLP domain papers. However, most examples in QASPER represent shallow questions focused on completing concepts because the annotators produced the questions after reading only the title and abstract of a provided paper. Additionally, in QASPER, more than 70% of answerable questions consist of short-form answers, such as yes/no and small extractive span. In contrast, we ask our annotators to read further into main sections, demanding various types of questions based on our studied schema. As a result, the questions in our

QASA cannot be simply answered with extracting spans form selected evidence paragraph. It truly urges full-stack reasoning.

Open-domain Long-form QA ELI5 (Fan et al., 2019) collected open-domain questions with paragraph-level answers collected from Reddit forum and extracted the relevant sentences from web documents, which are provided as supporting evidence. However, only 65% of the questions have sufficient supporting evidence, while all the answers of our dataset have evidence paragraph through associative selection. Stelmakh et al. (2022) have claimed that factoid questions in the ELI5 dataset are mostly ambiguous, and thus can be decomposed into sub-questions. ASQA (Stelmakh et al., 2022) requires to answer all the sub-questions over multiple passages. Our QASA differs in two ways: (1) our dataset includes evidential rationales, compositional element of long-form answer, (2) ours is made through a systematic composition that considers implicit relations between multiple rationales, rather than simply synthesizing the sub-answers.

Query-focused Multi-Document Summarization (qMDS) For qMDS, some datasets in various domains have been proposed, such as QMSum for meeting transcripts (Zhong et al., 2021), Squality for science fiction (Wang et al., 2022a), and AQuaMuSe for wikipedia (Kulkarni et al., 2020). The goal of these tasks is to find an answer over multiple documents, which is similar with ours. However, qMDS datasets such as AQuaMuSe and QMSum have the limitation of using noisy and insufficient contexts as multi-documents, since they used automatically-generated passages extracted by lexical matching. To address the issue of insufficiency of dedicated training data, the previous work (Baumel et al., 2018) adopts transfer learning techniques. In comparison to qMDS, our task provides human-annotated evidences aligned with a particular paragraph and answer summaries composed of multi-evidences. Additionally, qMDS focuses on summarizing text without redundancy, while we aim to generate rich long-form answers including multiple rationales.

3. Proposed Task

In this section, we propose a new task for question answering over scientific articles. The core idea of our proposed task is to answer the questions based on multiple evidence snippets that are spread over a long research paper. Specifically, we denote a question as q, an answer as a, and paragraphs in the paper as $P = \{p_1, ...p_N\}$. A one-step approach to process N paragraphs would be adopting length-scalable transformer such as LongFormer (Beltagy et al., 2020), which enables to encode the multiple snippets at once. In contrast, our advanced questions triggered from research papers requires to connect between rationales for deep reasoning. Hence, we design this problem as multi-step subtasks: (1) *associative selection*, (2) *evidential rationale-generation*, and (3) *systematic composition*. Figure 2 shows the overview of our approach.

Associative Selection While research papers have multiple paragraphs (*e.g.*, 20-60 paragraphs), the first step is to extract associative knowledge from the paragraphs, corresponding to a question. Specifically, given question q and paragraphs $P = \{p_1, ...p_N\}$, we aim to select evidential paragraphs $\bar{P} = \{\bar{p}_1, ...\bar{p}_k\}$ that contains an answer or rationales to question q, where $k \ll N$. While the previous work (Rajpurkar et al., 2018) aims to classify answerability whether a given passage contains answer a to question q, the answer in our task is composed of multiple rationales including a main answer. Our associative selection task can be viewed as the super-task of answerability (*i.e.*, answerable is evidential, but not the reverse).

Evidential Rationale-Generation In this step, we generate an evidential rationale on each selected paragraph, which could be part of a final long-form answer in the next step. Based on the prior work about discourse structure of answers to complex questions (Xu et al., 2022), the evidential rationale can be the (1) main answer (*i.e.*, main content of the answer which directly addresses the question), (2) elaboration (*i.e.*, sentences which elaborate on the main answer), and (3) auxiliary information (*i.e.*, background knowledge that could be helpful to the user). Specifically, we denote the evidential rationale that is inferred from (q,p_i) as e_i. That is, from the selected $\bar{P} = \{\bar{p}_1, ...\bar{p}_k\}$, we obtain a list of evidential rationales $\{e_1, ..., e_k\}$.

Systematic Composition To provide concise and readable information to users, the goal of this last step is to systematically compose all the evidential rationales $\{e_1, ..., e_k\}$ into a final comprehensive answer a. Assuming that the answer is composed of multi-rationales, we aim to preserve all the rich rationales in the final answer, except duplicated texts. Specifically, we aggregate the list of texts $\{e_1, ..., e_k\}$ into a single context, and then compose final answer a from the context. The answer a grounded on a given paper could be viewed as comprehensive explanations about question q.

4. Building the QASA Dataset

Prior to data collection, we conducted a preliminary study for identifying what kinds of questions are raised when reading papers. Based on our findings, we design a schema to collect diverse and balanced questions with different levels of reasoning. As the source of the QASA, we gather a set of open-access AI/ML papers. To collect advanced questions that require reasoning over evidential rationales, we recruited AI/ML practitioners or researchers who regularly

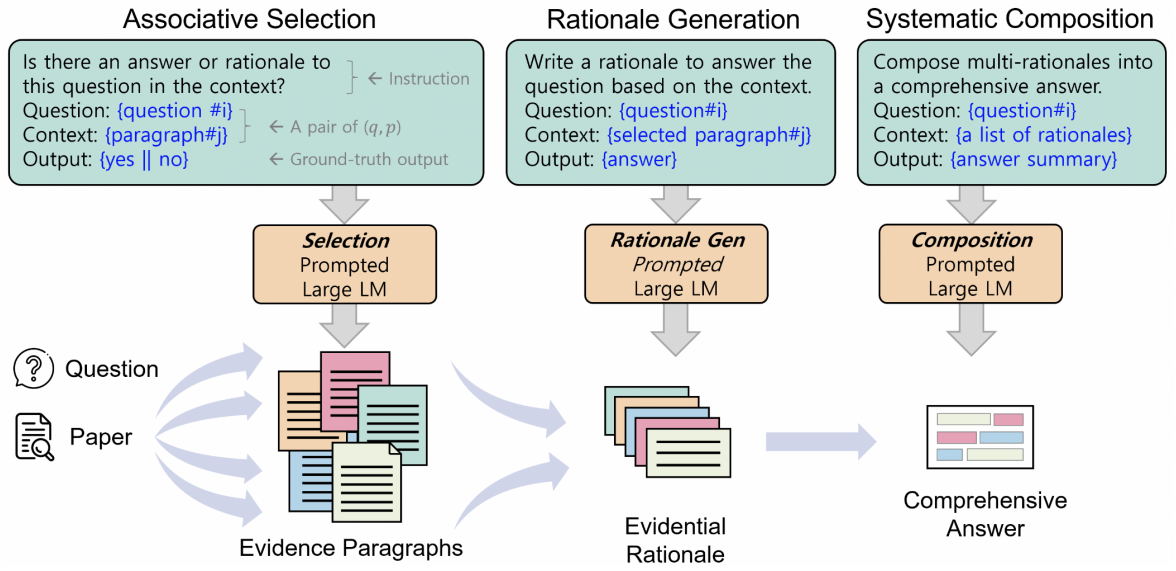

Figure 2. An overview of QASA approach. The language model works depending on task-specific instructions.

read research papers and conducted two separate sessions from the perspective of both *readers* and *authors*.

4.1. Preliminary Study

In the aim of identifying what kinds of questions readers ask while reading, we conduct a think-aloud study ($N = 10$), a standard approach in human-computer interaction (HCI) for capturing human's intent during a task. Our analysis of 127 questions revealed that 67% of questions required a two-stage process, and the types of reasoning needed in the second stage varied even among these questions. Referring to the prior literature on question taxonomy in the education domain, there are distinct types of questions ranging from surface questions to deeper questions that require more reasoning and interpretation to answer (Graesser et al., 1992; Graesser & Person, 1994). To gather diverse and balanced types of questions, we design a schema for paper questions by adapting the prior literature in education to a paper reading context and interpreting data collected from our think-aloud study. This schema includes not only questions requiring second stage reasoning, but also a spectrum of reasoning types needed to answer them. The definitions of each question type are shown below and detailed explanations with examples can be found in Appendix B.

- **Surface questions** aim to verify and understand basic concepts in the content. The answer content is directly related to the words in the question and immediate context. This type includes *verification, distinctive, concept completion* questions.
- **Testing questions** are focused on meaning-making and forming alignment with readers' prior knowledge. These questions aim to find similar examples (*example*), quantify variables (*quantification*), and find meaning and make comparisons across concepts (*comparison*).
- **Deep questions** ask about the connections among the concepts in the content and elicit advanced reasoning in logical, causal, or goal-oriented systems. This type includes *causal antecedent, causal consequence, goal orientation, instrumental/procedural, rationale, expectation* questions.

4.2. Papers

To collect papers, we adopt S2ORC (Lo et al., 2020), a collection of machine-readable full text for open-access papers, and the arXiv[1] paper collection. We only use papers within the CS.AI domain in the arXiv dataset and apply two filtering criteria to the papers in the S2ORC collection: (1) published after 2015 and (2) has more than 100 citations.

4.3. Data Collection

With the aim of collecting various advanced questions (surface to deep), we conduct two types of sessions, reader sessions where we collected QAs from general readers and author sessions where authors annotated questions about their own papers. We perform author sessions since authors are the optimal annotators who can make challenging and insightful questions that could be asked by experts, like reviewers—granting greater diversity to the questions in our benchmark. For the reader session, to make the data collection process similar to a real context, we decouple the

[1] https://arxiv.org/help/bulk_data

questioning and answering phase following the collection process of QASPER (Dasigi et al., 2021). For both tasks, we recruited graduate students studying AI/ML and freelancers practicing AI/ML through professional networks and Upwork.[2] For the answering task, we qualify annotators through the exams related to our task and experience in the domain. Details on the data collection procedure and workers' information are given in Appendix A.

Questions To ensure that our questions are realistic, we allow annotators to choose papers that they wanted to read. Additionally to replicate differing reading styles, we asked annotators to follow one of two scenarios: read all the sections in the paper (*i.e.*, deep reading) or read only certain sections (*i.e.*, skim reading). To collect diverse types of questions, we provide them with the question schema and asked them to make a balanced number of questions for each type. In the same vein, we also recommend annotators to make at least one question per subsection that they read. When annotating questions, annotators were instructed to write the trigger sentences that raised the question but that did not contain the answer. While they were not used in this work, trigger sentences could be used in future research for question generation from long-form text and to complement the ambiguity of questions that occur in long-form text.

Answers To collect answers, we ask answerers to choose papers from the papers that the questioners worked on. We guide answerers to compose their answers into a comprehensive passage based on the their own-generated evidential rationales from the selected paragraphs. To let them follow our guideline more easily, we provide the annotators with the answering interface when answering the questions. They were shown the question, the full paper, the name of the section that triggered this question, and ten paragraphs that are the most relevant to the question. We provide top ten paragraphs by following that existing IR research (Carterette et al., 2010) adopted a pooling method, where top ranked documents are selected to create the pool of documents that need to be judged when creating evaluation dataset. Our top-10 relevant paragraphs were chosen with an off-the-shelf embedding model.[3] When answering each question, annotators were asked to do the following subtasks.

First, they are asked to look through the ten relevant paragraphs and, for each, make a binary decision as to whether the paragraph is evidence paragraph. If there is no relevant paragraph chosen to have evidence, annotators could freely choose other paragraphs from the paper as having evidence in addition to the ten paragraphs we provided. Second, for each paragraph that was chosen, annotators are instructed to write evidential rationale from that paragraph. Evidential rationale could be the (1) main answer to the question, (2) elaboration, or (3) auxiliary information (Xu et al., 2022). Third, they write a final comprehensive answer by composing the multiple evidential rationales that they generated for each evidence paragraph. When the answer cannot be fully answered even after composing multiple evidential rationales, annotators are instructed to answer as much as possible with the available information and then specify which part of the question cannot be answered. When a question is completely unanswerable, we ask annotators to indicate that the author do not provide an explanation for the missing information and to specify what information is missing. Finally, they annotate whether writing the final answer requires to compose multiple evidential rationales (True) or not (False)—*i.e.*, no complex reasoning is needed and they only simplify text from the paper without adding redundancies.

All annotators who ask questions and write answers conducted a practice session to familiarize themselves with the annotation guidelines. Annotations from the practice sessions were reviewed by two authors, and discrepancies between the annotators and the guidelines were discussed. Additional practice sessions were conducted for annotators with substantial discrepancies. If annotators were judged as not having sufficient background knowledge or understanding of the task even after these sessions, we did not let them participate in the tasks.

Authors We recruited paper authors to annotate QAs for their own papers to cover deeper questions that existing datasets rarely cover. We instructed authors to make only **testing** and **deep** types of questions and to annotate trigger sentences that might cause readers to become curious about that question. The rest of the annotation process is similar to the readers' sessions. We recruited 17 authors whose domains are distributed in CV, NLP, GNN, generative models, and music information retrieval.

4.4. QASA Analysis

Representative examples from QASA is in Table 7 (in Appendix E).

Question types In terms of question types, two domain experts manually evaluated 100 randomly sampled questions. 89% of the annotated question types were aligned with domain experts' annotations.[4] To describe the diversity of our dataset, we analyze the distribution of the types of the questions in QASA. Among the three types, 39.4% of the questions are deep questions, 30.0% are testing, and 30.7% are surface-level. Among the deep questions, instrumental

[2] https://upwork.com/

[3] https://api.openai.com/v1/embeddings

[4] Two domain-experts independently judged these and achieved Cohen's κ scores of 0.91.

sub-type (12%) accounts for most of the deep questions, and comparison (11%) and concept completion (17%) are the most annotated questions for the testing and surface questions, respectively.

Distribution of evidential rationale We also analyze to identify the number of evidential rationales that are needed to answer the questions depending on their types. Among all the questions, 12% of questions are annotated as having no evidential rationales, which means that they are unanswerable questions. Out of the answerable questions, the average number of evidential rationales is 1.67. Surface questions need the most evidential rationales (1.73) while testing questions and deep questions need 1.66 and 1.63, respectively, which implies that our surface-level questions like *"Do the authors claim that bigger datasets would improve the performance and expressiveness of reading comprehension models?"* also need systematic reasoning to answer.

Composition, Correctness, Groundedness On average, 49.6% of answers require annotators to compose evidential rationales, while the rest (50.4% of answers) only need simplifying redundant rationales. To analyze which question type requires the most reasoning to answer, we analyze the ratio of compositionality depending on the question type. Deep questions need composing the most (44.6%) in comparison with testing (29.0%) and shallow (26.4%) questions. To estimate the correctness of the answer annotations and groundedness of the answer annotation, domain experts manually analyzed 100 randomly sampled questions. We find that 90% of the answers are correct and 87% are grounded well on the paper.

5. QASA Approach

In this section, we propose a QA approach for QASA over research papers. Our task requires to answer questions based on multiple passages whose supporting evidences are spread over a whole paper. As above-mentioned, our tasks consist of (1) *associative selection*, (2) *evidential rationale-generation*, and (3) *systematic composition*. As shown in Figure 2, we train LM models with multi-task instructions, following recent works (Chung et al., 2022; Wei et al., 2021; Aribandi et al., 2021; Sanh et al., 2021).

5.1. Multi-step QA system based on LM

Pre-processing via Retrieval Before the first step of *associative selection*, we consider pre-processing step using a retrieval model to narrow the search space, from a whole paper to top-N related paragraphs (we set N=10). This enables the efficient selection step, while compromising the recall of evidential paragraphs. Specifically, we used the off-the-shelf model provided by OpenAI, and leave the question of improving retrieval for future work. Through the retrieval, we encode all paragraphs in the given paper and a target question into dense vectors, and extract top-N nearest neighbor paragraphs by using cosine similarity.

Finetuning Large Language Model with Multitask Instructions We finetune large language models (LLMs) on a mixture of our subtasks through instruction tuning (Wei et al., 2021). As in previous work (Wei et al., 2021; Aribandi et al., 2021), it is known that instruction tuning makes LMs generalizable on unseen tasks. As shown in Figure 2, a single LM takes task input with instruction that indicates each subtask. The output of the previous step is sequentially passed to the next step. However, in the *selection* task, if the model does not select any paragraph as evidence, it also cannot generate rationales or answers. Instead of solving this problem, we used top-3 paragraphs if none were selected, which is left to future work. For task-specific prompts, we used manually-written instructions for each subtask (See Appendix D). As state-of-the-art LLMs, we consider the following models:

- T5 (Raffel et al., 2020) (Version 1.1, LM-Adapted): it is pretrained on Common Crawl (Raffel et al., 2020) using Transformer with encoder-decoder architecture.
- T0 (Sanh et al., 2021): starting from T5, it is further trained on 8 downstream tasks.
- FLAN-T5 (Chung et al., 2022): similar with T0, it is further trained with scaling up multi-tasks (1k+) including reasoning tasks.
- GALACTICA (Taylor et al., 2022): it is pretrained on a large collection of scientific papers, with the decode-only architecture like GPT.

5.2. Training Data

No training resources have been proposed that support our full-stack QA, and we therefore exploit public and synthetic data for the purpose of each subtask. Table 2 shows a summary of used public data.

Task	Dataset
Associative Selection	QASPER, ASQA
Rationale Generation	QASPER
Answer Composition	ASQA, ELI5

Table 2. Training Resources for our QA system.

For *associative selection*, we adopt answerability labels – whether the pair of question and knowledge is answerable or not. In case of ASQA (Stelmakh et al., 2022), we treat the pair of (q and p^+) as a positive example, and (question q,

randomly-sampled p^-) as negative. For QASPER (Dasigi et al., 2021), we leverage pairs of (question q, gold paragraph p^+ that contains an answer). The limitation of adopting these datasets is that they aim to capture the presence of an answer, while we target that of evidential rationales, which may affect recall of rationales.

The *rationale-generation* task requires to generate evidential rationale e from (question q, paragraph p). Unfortunately, to the best of our knowledge, there is no data to support this task. As an alternative source, we used the triplets of (q,p,a) in QASPER (Dasigi et al., 2021), where gold knowledge p always contains information about answer a to q. We treat answer a as evidential rationale, since the QA labels in QASPER do not require to reason over multiple passages, which is expected to learn the ability of extracting question-focused evidences from a given paragraph.

Lastly, for *systematic composition*, we adopt long-form QA data with multiple evidences. We select ELI5 and the subset of ASQA, which provide selected evidence passages from the pool of passages for answer generation. That is, this task is to generate answer a inferred from the context (question q, multi-evidences $\{e_1, ...e_n\}$), which requires to consolidate and summarize scattered information.

For synthetic data, recent works distill training data from InstructGPT (Wang et al., 2022b; Honovich et al., 2022). Inspired, we distil training examples for QASA from large language models by prompting instruction and an in-context example. We use OpenAI's InstructGPT (text-davinci-003) with the temperature set to 0.1, which is the state-of-the-art model on many NLP tasks. Specifically, we first extract AI and ML papers from arXiv, and generate questions over each paragraph sampled from the papers. Then, given the questions, we test InstructGPT following instructions in our subtasks, as shown in Appendix D. While LLMs have a general problem of factual inconsistency, known as *hallucination*, we found that InstructGPT performs well on the rationale generation task (See Table 3). Although there is no public data to support rationale generation, we can alleviate the insufficiency through evidential rationales obtained from InstructGPT, which would boost our full-stack QA.

6. Experiment

In this section, we evaluate our QASA approach on the proposed benchmark. In our experiment, we apply state-of-the-art LMs as two variants: Pretrained and Finetuned versions. In Sec 6.1 and 6.2, we automatically evaluate models on three subtasks: (1) *associative selection*, (2) *rationale-generation*, (3) *answer composition*, and their full-stack QA task. To complement automatic evaluations in our generation task, we conduct human evaluation in Sec 6.3 and an error analysis in Sec 6.4.

6.1. Experimental Setting

Evaluation of Subtasks and Full-stack QA For subtask evaluation, we provide oracle (or gold) contexts, in order to evaluate each subtask independently. In the *associative selection* task, we consider both positive paragraphs labeled by humans and negative paragraphs among top-10 retrieved results as candidate pool. For *rationale-generation*, we generate evidential rationale conditioned only on each of gold positive paragraphs. Similarly, for *answer composition*, we provide a list of gold evidential rationales as contexts. In contrast, for the full-stack QA, we consider the results of previous task as input to the next task sequentially, which could propagate the errors of the previous steps. Meanwhile, we conduct an ablation experiment, to directly generate final answers from selected paragraphs without *rationale-generation* ("w/o Rationale Gen" in Table 4).

Metric For *associative selection*, we measured the precision (P), recall (R), and F1 score. For *rationale-generation* and *answer composition* tasks, we used a standard text generation metric – ROUGE scores (Lin, 2004).

6.2. Main Results

Table 3 shows the automatic evaluation results of several QA systems on three subtasks and full-stack QA task.

Which pretrained LM is best? Among the pretrained LMs, INSTRUCTGPT (175B) outperformed others. Especially in the *rationale-generation* task, it shows the best performance among all models. Among T5-based LMs, the number of downstream tasks used during training had a significant impact on the performances in full-stack QA, showing T5 < T0 < FLAN-T5.

Which finetuned LM is best? When comparing finetuned T0, T5, and FLAN-T5, these models show little difference in performances on three subtasks. However, FLAN-T5 outperformed all other LMs on the full-stack QA, even the state-of-the-art model, INSTRUCTGPT (175B). Based on this observation, we suggest the finetuned FLAN-T5 could serve as a good test-bed for QASA.

The effect of training resources we curated For an ablation study, we trained individual Flan-T5 on each one of four datasets (QASPER, ASQA, ELI5, Augmented Data from GPT (or GPT AUG)). Through the comparison, we can observe negative transfer across datasets, *e.g.*, FLAN-T5 trained on ASQA-ONLY shows the best results in the *answer composition* task, outperforming FLAN-T5 trained on combined data. Meanwhile, training of GPT AUG improved significantly the performances in the *rationale-generation* task, which is essential for our full-stack QA, as other resources do not contain rationales.

Method	Associative Selection			Rationale Generation			Answer Composition		
	(P)	(R)	(F1)	(R-1)	(R-2)	(R-L)	(R-1)	(R-2)	(R-L)
Pretrained LMs (Accessible Checkpoints or API)									
GALACTICA(6.7B)	18.70	<u>94.28</u>	29.36	7.06	0.50	5.01	8.93	1.03	6.84
T5 (3B)	6.85	6.83	5.78	26.99	11.31	20.64	27.46	16.08	21.85
T0 (3B)	6.92	7.60	6.39	20.19	9.75	17.71	32.75	20.49	29.30
FLAN-T5 (3B)	37.50	38.57	34.64	20.30	11.62	18.36	40.90	27.30	35.78
INSTRUCTGPT (175B)	31.78	51.97	34.72	**41.27**	**24.69**	**33.64**	47.27	28.22	36.09
Finetuned LMs (on Collected Data)									
GALACTICA (6.7B)	31.70	47.39	33.32	8.45	1.07	6.98	13.90	2.44	10.41
T5 (3B)	<u>39.79</u>	56.56	40.71	26.73	13.02	22.64	46.40	29.60	38.55
T0 (3B)	39.04	77.29	<u>45.16</u>	27.86	13.40	23.45	46.78	29.29	38.24
FLAN-T5 (3B)	37.13	59.97	**45.86**	27.63	13.65	23.33	45.59	28.80	37.24
FLAN-T5 (3B, QASPER-ONLY)	30.67	91.43	41.67	23.47	13.13	20.94	40.51	26.88	35.57
FLAN-T5 (3B, ASQA-ONLY)	21.39	**98.97**	32.84	24.72	<u>13.90</u>	21.62	**48.97**	**31.93**	**40.34**
FLAN-T5 (3B, ELI5-ONLY)	**46.80**	36.16	40.80	21.03	11.59	18.74	33.24	18.17	28.59
FLAN-T5 (3B, GPT AUG-ONLY)	31.58	28.82	27.85	<u>29.22</u>	13.69	<u>24.15</u>	<u>48.53</u>	<u>31.19</u>	<u>38.80</u>

Table 3. The results of baseline systems on three subtasks in QASA, measured by Precision, Recall, F1 score, and ROUGE scores. The best results in each column are **bold**-faced, and 2nd best results are <u>underlined</u>.

Method	Full-stack QA		
	(R-1)	(R-2)	(R-L)
Pretrained LMs (Accessible Checkpoints or API)			
GALACTICA (6.7B)	15.56	3.65	11.44
T5 (3B)	9.83	0.58	8.01
T0 (3B)	15.60	4.28	12.15
FLAN-T5 (3B)	22.48	9.52	18.45
INSTRUCTGPT (175B)	27.11	11.90	19.75
Finetuned LMs (on Collected Data)			
GALACTICA (6.7B)	20.93	6.16	15.01
T5 (3B)	26.66	11.45	20.73
T0 (3B)	<u>29.75</u>	<u>13.13</u>	<u>22.75</u>
FLAN-T5 (3B)	**32.22**	**14.62**	**24.53**
w/o Rationale Gen	27.73	11.31	19.32

Table 4. The results of full-stack QA systems on QASA.

Does our task indeed need rationale-generation? For our full-stack QA, while we first generate rationales and then compose them into a final answer, we can directly generate an answer from selected paragraphs, skipping the step of rationale generation. However, as shown in Table 4, FLAN-T5 "w/o Rationale Gen" showed poor performance, compared to our three-step approach, which means the rationale generation step is crucial for the full-stack QA.

The failure of Galactica Although GALACTICA was pretrained on a large-scale collection of research papers, it performed worse on overall tasks compared to other models. The low performance of GALACTICA was consistently observed in Singhal et al. (2022), compared to PubMedGPT of 2.7B. We empirically found that GALACTICA often answered either "yes" or "no", and terminated the generation, in which case the Rouge score is almost zero.

6.3. Human Evaluation

Although automated metrics can measure crucial aspects of our task, they are not guaranteed to closely approximate the judgment of humans, whose satisfaction is an overarching goal of a QA system. Therefore, we performed human evaluations based on the dimensions that should be satisfied in this task.

We conducted a pairwise evaluation scheme where evaluators compare two answers to the same question, inspired by Stelmakh et al. (2022). We provided two responses to each human evaluator, one from ours and the other from InstructGPT. The human evaluators could read the rationales and the generated responses side-by-side. Then, the evaluators were asked to choose the better answer in terms of four criteria: Groundedness, Completeness, Specificity, and Fluency, following prior work (Stelmakh et al., 2022; Thoppilan et al., 2022). For each data point, we assigned three evaluators to collect three trials of such pairwise judgments. The scoring system awards one point for a win and half a point for a tie in pairwise comparisons. The annotations were collected on 100 QA pairs by 9 experts.

The results of this human evaluation in Figure 3 show that the answers from our full-stack QA tend to be more complete and grounded than those from InstructGPT, which is consistent with the results from the automatic evaluation. In

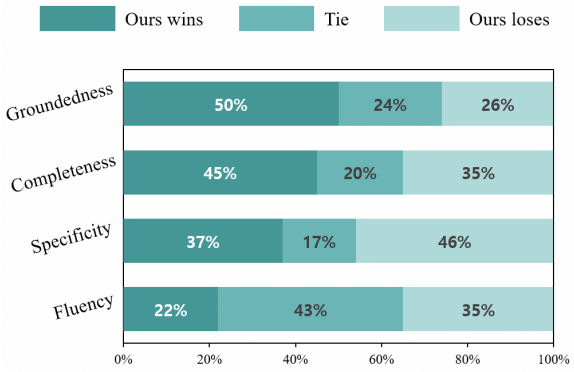

Figure 3. The results of human evaluation, comparing Ours to InstructGPT on four dimensions across 100 samples.

contrast, the InstructGPT's answers tend to be more fluent and specific, regardless of the reliability of its generated text. We also added some qualitative examples to show how the answers generated by our approach differ to those by InstructGPT in Appendix F.

6.4. Error Analysis

To gain a deeper understanding of the model's errors, we sample 50 test examples with Rouge-L scores below 10 (*i.e.*, bottom 25%). We exclude instances that are unanswerable based on the given paper. We then classify errors into five categories, ranging from E1 to E5.

E1 refers to cases where the model incorrectly classified the question as unanswerable. E2 is the generation of irrelevant content. E3 is cases where the model provides implicit evidence but fails to generate an explicit answer. E4 refers to cases where the generation is not factually grounded on the source document. Lastly, E5 refers to cases with low completeness, where the generation only covers a partial answer (i.e., a sub-question). Additionally, a low Rouge score does not necessarily indicate a wrong generation. We identify two correct scenarios for this (C1 and C2). C1 refers to cases where the human labels are incorrect. C2 is cases where both the generation and human label are correct, but the lexical overlap between the two texts is low due to the diversity of expressions.

Table 5 shows error analysis results. 36% of InstructGPT's answers and 34% of ours belong to C1 and C2: cases with low ROUGE score, but correct. 48% of InstructGPT's answers are cases of refusal to answer (*e.g.*, "I cannot find any specific information..."), although the context contains relevant evidences. We conjecture that InstructGPT has been trained to avoid answering in uncertain cases for safety. In contrast, our system did not generate such refusal responses, since there is no such example in our training data. 44% of our system's answers are irrelevant to a given question, although the text is grounded on evidence.

Type	Instruct GPT	Our Model
C1: incorrect human label	10%	10%
C2: low lexical overlaps	26%	24%
E1: predict unanswerable	48%	0%
E2: irrelevant generation	8%	44%
E3: failure of answering explicitly	0%	8%
E4: failure of grounding	6%	6%
E5: low completeness	2%	8%

Table 5. Error Analysis of InstructGPT and Ours.

7. Limitation

While we proposed a new benchmark for QA task on scientific articles, evaluation is becoming difficult, especially on recently emerging language models (InstructGPT as well as ChatGPT, Bard). Such language models aim not only for accurate responses, but also for longer responses through structured writing. Hence, evaluation metrics using string matching (such as ROUGE) may not represent the overall quality of generated results. The concurrent work showed that none of automatic metrics reliably matches human judgments of overall answer quality (Xu et al., 2023). Future work for our QA task could look deeper into adopting multi-faceted evaluations.

8. Conclusion

Conventional information search requires a series of non-trivial efforts from retrieving and reranking relevant information to manually reading and restructuring the selected information. Due to growing volumes of scientific papers and professional articles, the traditional process is no longer feasible, urging an innovation in knowledge processing and reasoning. Generative QA would be a promising alternative, but it lacks appropriate benchmark and principled methodologies that are focused on human intellectual capabilities: full-stack reasoning.

In this paper, we propose the QASA: a novel benchmark dataset and a computational approach. Our QASA benchmark guides expert readers and paper authors to generate various types of questions and answers from surface to testing and deep levels. Our QASA approach decomposes the full-stack reasoning process into three reasoning subtasks: associative selection, evidential rationale-generation, and systematic composition. By modeling each subtask by pre-trained LM, we show that FLAN-T5 finetuned on public and synthetic data could serve as the best test-bed for our QASA, proposing a new horizon of full-stack cognitive reasoning on scientific articles such as research papers and manuscripts.

References

Aribandi, V., Tay, Y., Schuster, T., Rao, J., Zheng, H. S., Mehta, S. V., Zhuang, H., Tran, V. Q., Bahri, D., Ni, J., et al. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*, 2021.

Baumel, T., Eyal, M., and Elhadad, M. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models, 2018. URL https://arxiv.org/abs/1801.07704.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Carterette, B., Kanoulas, E., and Yilmaz, E. Low cost evaluation in information retrieval. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M. A dataset of information-seeking questions and answers anchored in research papers. In *North American Chapter of the Association for Computational Linguistics*, 2021.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Evans, J. S. B. Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, 18(1):5–31, 2012.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, 2019.

Graesser, A. C. and Person, N. K. Question asking during tutoring. *American educational research journal*, 31(1): 104–137, 1994.

Graesser, A. C., Person, N., and Huber, J. Mechanisms that generate questions. *Questions and information systems*, 2:167–187, 1992.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938. PMLR, 2020.

Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022.

Izacard, G. and Grave, É. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Kulkarni, S., Chammas, S., Zhu, W., Sha, F., and Ie, E. Aquamuse: Automatically generating datasets for query-based multi-document summarization, 2020. URL https://arxiv.org/abs/2010.12694.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019a.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A. P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q. V., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019b.

Lee, M., He, X., Yih, W.-t., Gao, J., Deng, L., and Smolensky, P. Reasoning in vector space: An exploratory study of question answering. 2016.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, Y., Hashimoto, K., Zhou, Y., Yavuz, S., Xiong, C., and Philip, S. Y. Dense hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 188–200, 2021.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Pampari, A., Raghavan, P., Liang, J. J., and Peng, J. emrqa: A large corpus for question answering on electronic medical records. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Pappas, D., Androutsopoulos, I., and Papageorgiou, H. Bioread: A new dataset for biomedical reading comprehension. In *International Conference on Language Resources and Evaluation*, 2018.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Rajendran, J., Ganhotra, J., Singh, S., and Polymenakos, L. Learning end-to-end goal-oriented dialog with multiple answers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2021.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Stavropoulos, P., Pappas, D., Androutsopoulos, I., and McDonald, R. T. Biomrc: A dataset for biomedical machine reading comprehension. *ArXiv*, abs/2005.06376, 2020.

Stelmakh, I., Luan, Y., Dhingra, B., and Chang, M.-W. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*, 2022.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. Lamda: Language models for dialog applications, 2022.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

Tsujii, T. and Watanabe, S. Neural correlates of dual-task effect on belief-bias syllogistic reasoning: a near-infrared spectroscopy study. *Brain research*, 1287:118–125, 2009.

Wang, A., Pang, R. Y., Chen, A., Phang, J., and Bowman, S. R. Squality: Building a long-document summarization dataset the hard way, 2022a. URL https://arxiv.org/abs/2205.11465.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Wason, P. C. and Evans, J. S. B. Dual processes in reasoning? *Cognition*, 3(2):141–154, 1974.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., and Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Xu, F., Li, J. J., and Choi, E. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3556–3572, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.249. URL https://aclanthology.org/2022.acl-long.249.

Xu, F., Song, Y., Iyyer, M., and Choi, E. A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201*, 2023.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X., and Radev, D. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472.

A. Dataset collection details

Question writers and answer writers were paid US $28 and $63 (respectively) per paper on average and we have 26 question-makers and 28 answer-makers in terms of reader sessions and N authors in author sessions. We did not specify the number of questions per paper to allow annotators to create meaningful questions rather than be forced to add unnecessary questions. However, we recommend making around 15 questions per paper in order to guarantee dataset size.

Workers (*i.e.,* question/answer writers and authors) provided basic information about their expertise in AI/ML and question writers were asked to provide how familiar they already were with the paper for which they asked questions. The field of workers was in the order of CV, NLP, and Applied ML, and there were also workers from theoretical ML, GNN, RL, MLOps, music IR, and Human-centered AI. Most question writers (84.6%) had some experience in AI/ML, with 31.8% having more than four years of experience. Similarly, the majority of answer writers (88%) had experience in AI/ML, and 36.4% of them had over four years of experience. 50% of the authors have over four years of relevant experience and 66.7% of the authors have submitted three or more papers from their domains. 89% of the papers were seen by the question writers for the first time.

B. Question level taxonomy

For each question level, we provided the types of questions that are in that level and examples for each of these question types.

B.1. Surface questions

Surface-level questions aim to verify, compare, and understand basic concepts in the content. The answer content is directly related to the words in the question and immediate context.

Verification

- Is this true? Did an event occur?
- Examples
  - Did the authors have an experiment with training the state-of-the-art QA model with QuAC dataset?
  - They claim that LSTM can synthesize unseen compositions. Is this true?

Disjunctive

- Is X or Y the case?
- Examples
  - For metrics involving co-occurrence C, were they measured with the original C or the rectified C?

Concept completion

- Who? What? When? Where?
- Examples
  - What are the metrics used to measure the audio quality in the model comparison experiment?
  - Who were recruited as annotators of the entities and relations of concepts of lecture transcripts?

B.2. Testing questions

Testing questions are focused on meaning-making and alignment with readers' prior knowledge. The questions are marked by qualifying parameters of the components and generating initial interdependencies between the concepts. These questions aim to find similar examples, quantify the variables, find meaning and make comparisons across concepts.

Example

- What is an example label or instance of the category?
- Examples
  - What are the examples of the style of websites?

Quantification

- What is the value of a quantitative variable? How much? How many?
- Examples
  - How was the ratio of toxic words in the total vocabulary?
  - According to the statement that the validation set is 15% of tottal dataset, how many data points are in the validation set?

Definition

- What does X mean?
- Examples
  - What does "non-factoid" mean?
  - The result showed most dialogs in the QuAC dataset cover three to six of the chunks, but what does “chunk” mean?

Comparison

- How is X similar to Y? How is X different from Y?
- Examples
  - What points in DDPM are novel compared to LDM?
  - Likelihood-based methods do not suffer from the model-data mismatch issue. What are the benefits of using spectral methods instead of using standard probabilistic inference?

B.3. Deep questions

The questions ask connections among the concepts in the content and elicit advanced reasoning in logical, causal, or goal-oriented system.

Causal antecedent

- What state or event causally led to an event or state?
- Examples
  - Why would end-users want to stylize or customize websites?
  - Why do approaches that train transformation modules face difficulties in accessing prior knowledge with new concepts?

Causal consequence

- What are the consequences of an event or state? What if X occurred? What if X did not occur?
- Examples
  - The Low-rank Anchor Word algorithm (LAW) involves computing the QR decomposition of Y = QR. What is the additional cost incurred by this step?
  - The author used only 3-5 images of a user-provided concept to learn to represent it through new “words” in the embedding space, would results improve with more images? Why or why not?
  - While fine-tuning, the proposed method begins by unfreezing only the last layer and beginning training on that unfrozen layer only. Is this method likely to work for generative (encoder-only) models, or is this something that would work only in decoder-encoder models?

Goal orientation

- What are the motives or goals behind an agent’s action? Why did an agent do some action?
- Examples
  - Why was a large language model used in classifying the relation between concepts?
  - What are the different metrics are used in experiment 1 and 2?

Instrumental/procedural

- What plan or instrument allows an author to accomplish a goal? How did an author or author’s artifact do some action?

- Examples
  - How did the authors handled the issues with turker's different cultural backgrounds?
  - How does the proposed method address the issue of catastrophic forgetting?

Rationale

- How does the author show X (claim)? How does the result infer X (claim)? Why is it possible to say X (claim)?
- Examples
  - How do they show that single word embeddings capture unique and varied concepts?
  - How is "increased contextuality" observed in the data?

Expectation

- Why did some expected events not occur?
- Examples
  - Why the patterns over increasing x-axis values are not always consistent?

C. Training Details

All of our experiments were conducted using 16 A100 GPUs. To simplify all experiments, we fixed the initial learning rate to 1e-5. We trained all models until 5 epochs and selected the best checkpoint with average R-2 scores of *answer composition*on validation set.

D. Instructions

Our tasks consist of (1) *associative selection*, (2) *evidential rationale generation*, and (3) *systematic composition*. As shown in Table 6, we composed instructional templates for each task.

Table 6. Task-specific prompts for finetuning

Task	Templates

associative selection	`Is there a rationale or answer to this question in this context?`
	`Context: {paragraph}`
	`Question: {question}`

evidential rationale generation	`Write an answer or rationale to answer the question based on the context.`
	`Context: {selected paragraph}`
	`Question: {question}`
	`Answer:`

systematic composition	`Given the question, compose the multiple information into a coherent and comprehensive answer.`
	`Evidences: {e for e in evidences}`
	`Question: {question}`
	`Answer:`

E. Examples

Table 7 presents examples of QASA dataset. It includes the type of question, question, evidential rationales, answer, and composition.

F. Qualitative Examples of Answers

G. Ethics Statement

We present a new dataset that uses papers authored by other researchers. In compliance with copyright regulations, we have limited paper selection to papers available on arXiv that has been released under a Creative Commons Attribution license. Prior to conducting the annotation process, we obtained consent from the annotators once explaining the purpose the task. We collected information about annotators' background in AI/ML, their number of publications, and their familiarity with

Table 7. Examples from QASA.

Type	Question	Evidential rationales	Answer	Composition
Surface question	"Do the authors claim that bigger datasets would improve the performance and expressiveness of reading comprehension models?"	["Mentions that previous approaches in the literature that attempt to use synthetic dataset approaches (which can quickly be used to generated datasets of large sizes) have shown promise for reading comprehension tasks", "Authors mention that creating a large-scale labelled dataset as one of their contributions.", "Indicates that attention based models and LSTM models are effective at reading comprehension tasks.", "Explains that they (the authors) intend to provide a dataset that could be used to benchmark performance on reading comprehension tasks"]	"Based on the information in this paper alone, it is unclear if a bigger dataset would improve the performance of reading comprehension models. While authors explain that a key contribution they make is the creation of a real-world, massive labelled reading comprehension dataset, it is unclear if such a dataset is essential to improve the performance of reading comprehension models - the authors pitch their dataset-building approach also as a way of evaluating performance of these models, which is different from the dataset itself leading to better performance."	True
Surface question	"Which deep neural network architectures were used for experimental comparison of DeepFool algorithm with existing methods?"	["The experiments for the DeepFool method involve 6 different classifiers for 3 different datasets: MNIST: 2-layer FC, 2-layer LeNet; CIFAR-10: 3-layer LeNet, NIN; ILSVRC 2012: CaffeNet, GoogLeNet.", "The paper provides extensive experiments with 3 different datasets and 8 different classifiers."]	"Although the conclusion of the paper claims that 8 different classifiers were used, we can only see 6 classifiers with different datasets: 2-layer fully-connected network (MNIST), 2-layer LeNet (MNIST), 3-layer LeNet (CIFAR-10), NIN (CIFAR-10), CaffeNet (ILSVRC 2012), and GoogLeNet (ILSVRC 2012)."	False
Testing question	"What is the difference between BERT paper and RoBERTa paper's point of views? Give an answer in NSP loss and their performance perspective."	["In BERT paper, author said that removing NSP can hurt the performance of the model.", "In RoBERTa paper, author said that removing NSP improves downstream task performance."]	"In BERT paper, author said that removing NSP can hurt the performance of the model. However, in RoBERTa paper, author said that removing NSP improves downstream task performance. Therefore, point of views in terms of NSP is different between BERT and RoBERTa."	True
Testing question	"What does "interaction between the pixels to the text embedding through the diffusion process" mean?"	["During the diffusion process, we predict the noise of an image given a noisy image and text embedding using U shaped network. and this process yield our final image at the last step. and the interaction between the two modality occurs during the noise prediction. And the interaction between the visual and textual features are fused using Cross-attention layers that produce spatial attention maps for each textual token.", "Authors consider using internal cross-attention in their work, and the cross-attention maps are high-dimensional tensors that bind pixels and tokens extracted from the prompt text."]	"To answer this question we need to recall the diffusion process, which is in order to predict the noise of an image we have two inputs 1- noisy image and 2- text embedding, and the interaction between the two inputs are fused using Cross-attention layers that produce spatial attention maps for each textual token. and that is what is meant by the interaction between pixels to text embedding."	False
Deep question	"When defining the reading comprehension task, the authors explain that they wish to estimate p(alc, q). What would a model trained on this task do if the context 'c' itself had factually incorrect information?"	["This paragraph explains that the authors wish to estimate the conditional probability of an answer (a) being relevant to a question (q) given some context (c).", "Explains that the proposed dataset's objective is to test if a model is able to read and comprehend a document - not test a model's global knowledge of whether a statement is true or false.", "Mentions that the task they are building the dataset for is a reading comprehension task."]	"The authors are training a reading comprehension model. Therefore, if the context "c" has incorrect information, the model is likely to answer based on the factually incorrect information itself. The authors clearly explain that the task their model is being built for and evaluated on is of identifying answers from a given text (i.e. comprehension) and not knowledge of global correctness."	True
Deep question	"What weaknesses would a dataset that without entity replacement or anonymization have when training a reading comprehension model? Why is this a necessary step in the process?"	["Explains the difference between an original (unprocessed) data point and anonymized sample through an example. This paragraph points out that the non-anonymous version of the query could potentially be answered by an agent (either human or an ML model) even without reading the paragraph/context, while that would not be possible post-anonymization. This change ensures that the metric being measured is reading comprehension only and not anything else.", "Explains that they replace all entities with an abstract entity marker."]	"Since the authors are attempting to build a reading comprehension model, not anonymizing the entities before using the dataset might lead to a situation where models use external information, or statistics on the distribution/frequency of words themselves to guess answers. These steps are needed to ensure that models use the context to answer the questions."'	False

Table 8. The qualitative examples of answers, evidential paragraphs, and rationales from Ours and InstructGPT

Question	Answer-Ours	Answer-InstructGPT	Evidential Paragraphs	Rationales
Give two examples of public BERT-style english corpora.	*BookCorpus and English Wikipedia are two public BERT-style English corpora.*	*Two examples of public BERT-style English corpora are BookCorpus and OpenWebText. BookCorpus is a large-scale corpus of 11,038 books written in English, while OpenWebText is a collection of over 8 million webpages. Both corpora are used to train and evaluate natural language processing models, such as BERT.*	['We consider five English-language corpora of varying sizes and domains(...) We use the following text corpora:•BookCorpus (Zhu et al., 2015) plus English Wikipedia. This is the original data used to train BERT. (16GB).•CC-News, which we collected from the English portion of the CommonCrawl News dataset (Nagel et al.,, 2016). (...)]	['CC-News, BookCorpus and Wikipedia are BERT-style english corpora.']
How many tokens are changed to [MASK] in BERT training? Give a ratio.	*BERT training uses [MASK] to replace 80% of the tokens.*	*In BERT training, 15% of tokens are changed to [MASK], resulting in a ratio of 1:6.4 (15% of 512 tokens is approximately 80).*	['A random sample of the tokens in the input sequence is selected and replaced with the special token [MASK]. (...) BERT uniformly selects 15% of the input tokens for possible replacement. Of the selected tokens, 80% are replaced with [MASK], 10% are left unchanged, and 10% are replaced by a randomly selected vocabulary token.']	['Of the selected tokens 15%, 80% are replaced with [MASK] during training.']

the papers they are annotating. However, we did not collect personal identifiable information without the annotators' explicit consent, except for payment purposes. Additionally, the information was not included in the dataset we proposed.