A Appendix

A.1 Dataset documentation and intended uses

We follow datasheets for datasets guideline to document the followings.

A.1.1 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - QAConv is created to test understanding of informative conversations such as business emails, panel discussions, and work channels. It is designed for QA on informative conversations to fill the gap of common Wikipedia-based QA tasks.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - Salesforce AI Research team and HKUST CAiRE team work together to create this dataset.
- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
  - Salesforce AI research team funded the creation of the dataset.

A.1.2 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  - QAConv has conversations (text) among speakers (people) and a set of corresponding QA pairs (text).
- How many instances are there in total (of each type, if appropriate)?
  - QAConv has 34,204 QA pairs and 10,259 conversations. Each conversation has 568.8 words in average and the longest one has 19,917 words.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  - The conversations in QAConv are randomly sampled from several conversational datasets, including BC3, Enron, Court, Media, and Slack, and the number of samples is decided based on related work and the budget.
- What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  - Each sample has raw text of conversations, speaker names, and QA pairs.
- Is there a label or target associated with each instance? If so, please provide a description.
  - Each answerable sample has at least one possible answer in a list format.
- Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  - We do not include the crowd worker information due to the potential privacy issue.
- Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
  - N/A
- Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

  - Yes, we split training, development, and testing set by 80%, 10%, 10%. We split randomly within each data source.

- Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
  - There could have some potential noise of question or answer annotation.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions] (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
  - QAConv is self-contained.
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
  - No, all the samples in QAConv is public available.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  - No
- Does the dataset relate to people? If not, you may skip the remaining questions in this section.
  - Yes
- Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
  - QAConv contains different speakers with their names. Some samples have their role information, e.g., petitioner.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
  - Yes, because some of the conversations are coming from public forums, therefore, people may be able to find the original speaker if they find the original media source.
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual. orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
  - N/A.

A.1.3 Collection Process

- How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
  - The QA data is collected by Amazon Mechanical Turk. The data is directly observable.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
  - The QA data is collected by Amazon Mechanical Turk, we design a user interface with instructions on the top and then given partial conversation as context.
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

  - Crowdworkers. We paid them roughly $8-10 per hour, calculated by the average time to read and wriite one QA pair is approximately 4 minutes.
- Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
  - The data was collected during Feb 2021 to March 2021.
- Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
  - We have conduct an internal ethical review process by Salesforce ethical AI team, `https://einstein.ai/ethics`.
- Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
  - Yes.
- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
  - We obtain the data through AMT website.
- Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
  - Yes, the turkers know the data collect procedure. Screenshots are shown Figure 4, Figure 5, Figure 6 in the Appendix.
- Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
  - AMT has its own data policy. `https://www.mturk.com/acceptable-use-policy`.
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
  - `https://www.mturk.com/acceptable-use-policy`.
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
  - N/A

A.1.4 Preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the. remainder of the questions in this section.
  - We conduct data cleaning such as removing code snippets before asking the crowd workers to provide corresponding QA pairs. Thus, no additional cleaning or preprocessing is done for the released dataset, only the reading scripts used to change the format for model reading are used.
- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.
  - Yes, in the same link.
- Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
  - `https://github.com/salesforce/QAConv`

A.1.5 Uses

- Has the dataset been used for any tasks already? If so, please provide a description.
  - It is proposed to use for QA on conversations task.
- Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
  - It is a new dataset. We run existing state-of-the-art models and release the code at `https://github.com/salesforce/QAConv`
- What (other) tasks could the dataset be used for?
  - Many conversational AI related tasks can be applied or transferred, for examples, conversational retrieval and conversational machine reading.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
  - Different ways to disentangle conversations could impact the overall performance. In our current setting, we use and release the buffer-based chunking mechanism.
- Are there tasks for which the dataset should not be used? If so, please provide a description.
  - Conversations from Media corpus should not be used for commercial usage.

A.1.6 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
  - No.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
  - Release on Github. No DOI.
- When will the dataset be distributed?
  - It is released at `https://github.com/salesforce/QAConv`
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
  - BSD 3-Clause "New" or "Revised" License.
    `https://github.com/salesforce/QAConv/blob/master/LICENSE.txt`
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
  - No.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
  - Media dataset is restricted their conversations to be research-only usage.
    `https://github.com/zcgzcgzcg1/MediaSum`

A.1.7 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - Salesforce AI Research team. Chien-Sheng (Jason) Wu is the corresponding author.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  - Create an open issue on our Github repository or contact the authors (wu.jason@salesforce.com).
- Is there an erratum? If so, please provide a link or other access point.
  - No.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
  - No. If we plan to update in the future, we will indicate the information on our Github repository.
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
  - No.
- Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
  - Yes. If we plan to update the data, we will keep the original version available and then release the follow-up version, for example, QAConv-2.0
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
  - Yes, they can submit a Github pull request or contact us privately.

A.2 Accessibility

1. Links to access the dataset and its metadata.
   `https://github.com/salesforce/QAConv`
2. The data is saved in a json format, where an example is shown in the README.md file.
3. Salesforce AI Research team will maintain this dataset on the official company Github account.
4. BSD 3-Clause "New" or "Revised" License
   `https://github.com/salesforce/QAConv/blob/master/LICENSE.txt`

A.3 Data Usage

The authors bear all responsibility in case of violation of rights. We have used only the publicly available transcripts data and adhere to their guideline, for example, the Media data is for research-purpose only and cannot be used for commercial purpose. As conversations may have biased views, for example, specific political opinions from speakers, the transcripts and QA pairs will likely contain them. The content of the transcripts and summaries only reflect the views of the speakers, not the authors' point-of-views. We would like to remind our dataset users that there could have potential bias, toxicity, and subjective opinions in the selected conversations which may impact model training. Please view the content and data usage with discretion.

A.4 Test Data Additional Verification

After random split, we run an additional verification step on the test set. If the new collected answer is very similar with the original answer (FZR score > 90), we keep the original answer. If the new answer is similar within a margin (90 > FZR score > 75), we keep both answers. If the new answer is very different from the original answer (75 > FZR score), we will run one more verification step to get the 3rd answers. We pick the most similar two answers as the gold answers if their FZR score is > 75, otherwise, we manually looked into those controversial QA pairs and made the final judgement.

This process is only conducted in the test set and it could have multiple answers in the annotation (a list of string). In our released evaluation script, we take the maximal score from all potential answers to represent the result of each sample. One can observe a gap between our development set result and our test set results due to this additional verification step. Once the machine performance can surpass human performance, we will consider to release a harder holdout test set to evaluate model generalization.

A.5 License and Privacy

- BC3: Creative Commons Attribution-Share Alike 3.0 Unported License. (https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/bc3.html)
- Enron: Creative Commons Attribution 3.0 United States license. (https://enrondata.readthedocs.io/en/latest/data/edo-enron-email-pst-dataset/)
- Court: This material is based upon work supported in part by the National Science Foundation under grant IIS-0910664. Any opinions, findings, and conclusions or recommendations expressed above are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. (https://confluence.cornell.edu/display/llresearch/Supreme+Court+Dialogs+Corpus)
- Media: Only the publicly available transcripts data from the media sources are included. (https://github.com/zcgzcgzcg1/MediaSum/)
- Slack: Numerous public Slack chat channels (https://slack.com/) have recently become available that are focused on specific software engineering-related discussion topics (https://github.com/preethac/Software-related-Slack-Chats-with-Disentangled-Conversations)

A.6 Human evaluation description of human-written and machine-generated questions.

Rate [Fluency of the question]:

- (A) The question is fluent and has good grammar. I can understand clearly.
- (B) The question is somewhat fluent with some minor grammar errors. But it does not influence my reading.
- (C) The question is not fluent and has serious grammar error. I can hardly understand it.

Rate [Complexity of the question]:

- (A) The answer to the question is hard to find. I have to read the whole conversation back-and-forth more than one time.
- (B) The answer to the question is not that hard to find. I can find the answer by reading several sentences once.
- (C) The answer to the question is easy to find. I can find the answer by only reading only one sentence.

Rate [Confidence of the answer]:

- (A) I am confident that my answer is correct.
- (B) I am not confident that my answer is correct.

A.7 Computational Details

We run most of our experiments on 2 V100 NVIDIA GPUs with a batch size that maximizes their memory usage. We finetune T5-3B model on four A100 NVIDIA GPUs with several parallel tricks, such as fp16, sharded_ddp and deepseep. We train 10 epochs for t5 models and 5 epochs for BERT-based models. More training information is shown in `https://github.com/salesforce/QAConv`.

Table 9: Evaluation results: Molweni on the test set. * number is obtained from the original paper.

	Zero-Shot			Finetune		
	EM	**F1**	**FZ-R**	**EM**	**F1**	**FZ-R**
Human Performance	64.3	80.2	-	-	-	-
DialogueGCN*	-	-	-	45.7	61.0	-
DADgraph*	-	-	-	46.5	61.5	-
BERT-Large (SQuAD 2.0)	3626	45.90	56.90	53.43	66.85	73.50
RoBERTa-Large (SQuAD 2.0)	**38.42**	51.37	60.33	**53.92**	67.47	73.62
T5-Large (UnifiedQA)	34.52	**53.64**	63.08	52.14	69.04	**75.38**
T5-3B (UnifiedQA)	35.01	55.51	**64.14**	52.14	**69.21**	75.25

Table 10: Question type distributions: Top 10.

QAConv	**Squad 2.0**	**QuAC**	**CoQA**	**Molweni**	**FriendQA**	**DREAM**
what (29.09%)	what (49.07%)	what (35.67%)	what (31.02%)	what (65.9%)	what (19.97%)	what (53.33%)
which (27.21%)	how (9.54%)	did (19.19%)	who (13.43%)	how (11.4%)	who (18.1%)	how (11.32%)
how (11.54%)	who (8.36%)	how (8.13%)	how (9.38%)	who (7.54%)	where (16.07%)	where (10.29%)
who (9.99%)	when (6.2%)	was (6.05%)	did (8.0%)	why (5.57%)	why (15.99%)	why (7.94%)
when (6.03%)	in (4.35%)	are (5.45%)	where (6.41%)	where (5.54%)	how (15.14%)	when (5.05%)
where (4.48%)	where (3.62%)	when (5.43%)	was (4.53%)	when (1.84%)	when (11.76%)	who (2.89%)
why (2.75%)	which (2.83%)	who (4.62%)	when (3.29%)	which (1.53%)	which (0.51%)	which (2.84%)
in (1.79%)	the (2.47%)	why (3.11%)	why (2.73%)	whose (0.12%)	at (0.34%)	the (1.57%)
the (1.46%)	why (1.58%)	where (3.06%)	is (2.69%)	is (0.09%)	monica (0.34%)	according (0.59%)
on (0.38%)	along (0.36%)	is (1.74%)	does (2.09%)	did (0.08%)	whom (0.25%)	in (0.49%)
Other (5.27%)	Other (11.62%)	Other (7.55%)	Other (16.41%)	others (0.42%)	Other (1.52%)	Other (3.68%)

Table 11: Evaluation results: Full mode with DPR-wiki on the test set.

DPR-wiki	Zero-Shot			Fine-Tune		
	EM	**F1**	**FZ-R**	**EM**	**F1**	**FZ-R**
DistilBERT-Base (SQuAD 2.0)	28.23	31.61	47.24	37.23	44.33	56.84
BERT-Base (SQuAD 2.0)	25.21	28.84	45.24	37.71	45.04	57.56
BERT-Large (SQuAD 2.0)	**34.63**	**38.44**	**52.43**	41.05	47.56	59.73
RoBERTa-Base (SQuAD 2.0)	33.11	36.80	51.23	40.68	47.14	59.15
RoBERTa-Large (SQuAD 2.0)	33.63	37.35	51.62	**42.39**	**48.69**	**60.54**
T5-Base (UnifiedQA)	32.23	41.15	53.97	40.19	47.17	58.94
T5-Large (UnifiedQA)	36.37	44.86	56.89	42.19	49.04	61.01
T5-3B (UnifiedQA)	**38.22**	**46.53**	**58.44**	**43.56**	**49.89**	**61.43**

Table 12: Evaluation results: Chunk mode on the dev set.

	Zero-Shot			Finetune		
	EM	F1	FZ-R	EM	F1	FZ-R
DistilBERT-Base (SQuAD 2.0)	36.09	44.21	57.20	52.84	66.82	73.47
BERT-Base (SQuAD 2.0)	31.72	41.53	54.85	54.31	68.32	74.84
BERT-Large (SQuAD 2.0)	**47.04**	**58.14**	**67.29**	58.52	73.26	78.25
RoBERTa-Base (SQuAD 2.0)	43.91	54.30	64.66	57.94	72.24	77.59
RoBERTa-Large (SQuAD 2.0)	45.05	55.87	65.85	**61.45**	**75.15**	**80.13**
T5-Base (UnifiedQA)	46.51	61.83	68.93	59.17	73.24	78.27
T5-Large (UnifiedQA)	52.96	67.62	73.59	**61.39**	75.59	**80.14**
T5-3B (UnifiedQA)	**53.98**	**68.48**	**74.21**	61.31	**75.77**	80.08

Table 13: Retriever results: BM25 on the dev set.

	R@1	R@3	R@5	R@10
BM25	0.5835	0.7578	0.8037	0.8509

Table 14: Evaluation results: Full mode with BM25 on the dev set.

	Zero-Shot			Finetune		
	EM	F1	FZ-R	EM	F1	FZ-R
DistilBERT-Base (SQuAD 2.0)	26.60	32.35	48.65	36.41	47.20	58.98
BERT-Base (SQuAD 2.0)	22.76	29.49	46.14	37.49	48.36	59.85
BERT-Large (SQuAD 2.0)	**32.95**	**40.81**	**54.63**	40.42	51.74	62.21
RoBERTa-Base (SQuAD 2.0)	31.37	38.87	53.42	39.60	50.76	61.46
RoBERTa-Large (SQuAD 2.0)	31.87	39.24	53.70	**42.24**	**52.87**	**63.50**
T5-Base (UnifiedQA)	25.66	25.78	47.93	40.48	51.56	61.32
T5-Large (UnifiedQA)	32.51	43.61	54.79	41.68	52.88	62.19
T5-3B (UnifiedQA)	**34.53**	**46.55**	**56.73**	**42.33**	**53.94**	**63.10**

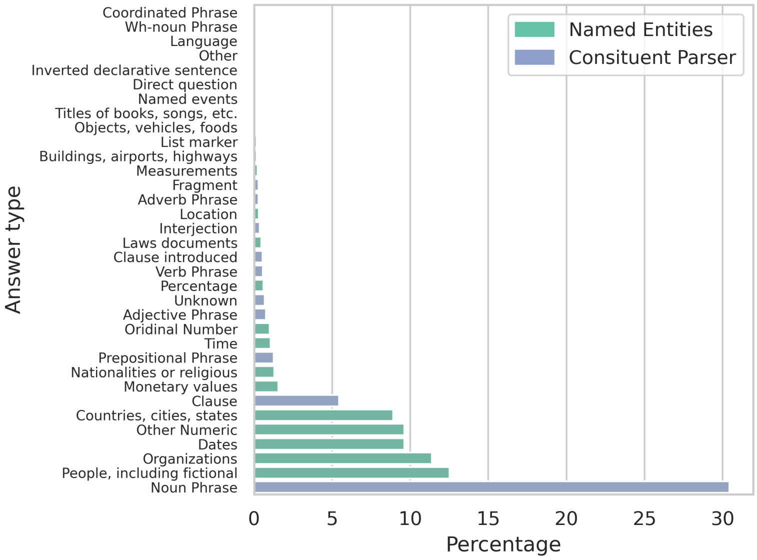
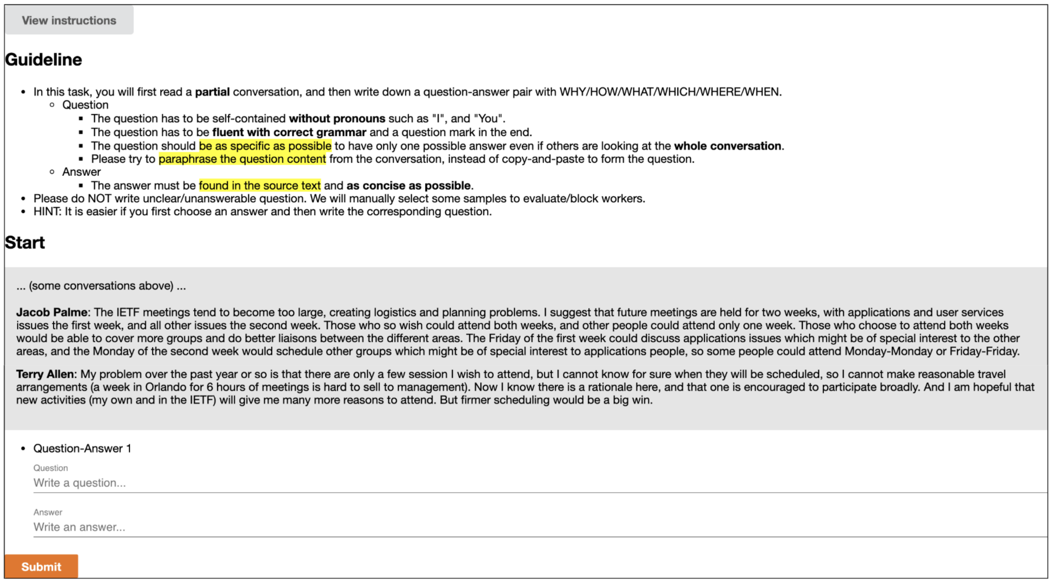


Figure 3: Diversity in answers in all categories.



Guideline

- In this task, you will first read a **partial** conversation, and then write down a question-answer pair with WHY/HOW/WHAT/WHICH/WHERE/WHEN.
  - Question
    - The question has to be self-contained **without pronouns** such as "I", and "You".
    - The question has to be **fluent with correct grammar** and a question mark in the end.
    - The question should be as specific as possible to have only one possible answer even if others are looking at the **whole conversation**.
    - Please try to paraphrase the question content from the conversation, instead of copy-and-paste to form the question.
  - Answer
    - The answer must be found in the source text and **as concise as possible**.
- Please do NOT write unclear/unanswerable question. We will manually select some samples to evaluate/block workers.
- HINT: It is easier if you first choose an answer and then write the corresponding question.

Start

... (some conversations above) ...

Jacob Palme: The IETF meetings tend to become too large, creating logistics and planning problems. I suggest that future meetings are held for two weeks, with applications and user services issues the first week, and all other issues the second week. Those who so wish could attend both weeks, and other people could attend only one week. Those who choose to attend both weeks would be able to cover more groups and do better liaisons between the different areas. The Friday of the first week could discuss applications issues which might be of special interest to the other areas, and the Monday of the second week would schedule other groups which might be of special interest to applications people, so some people could attend Monday-Monday or Friday-Friday.

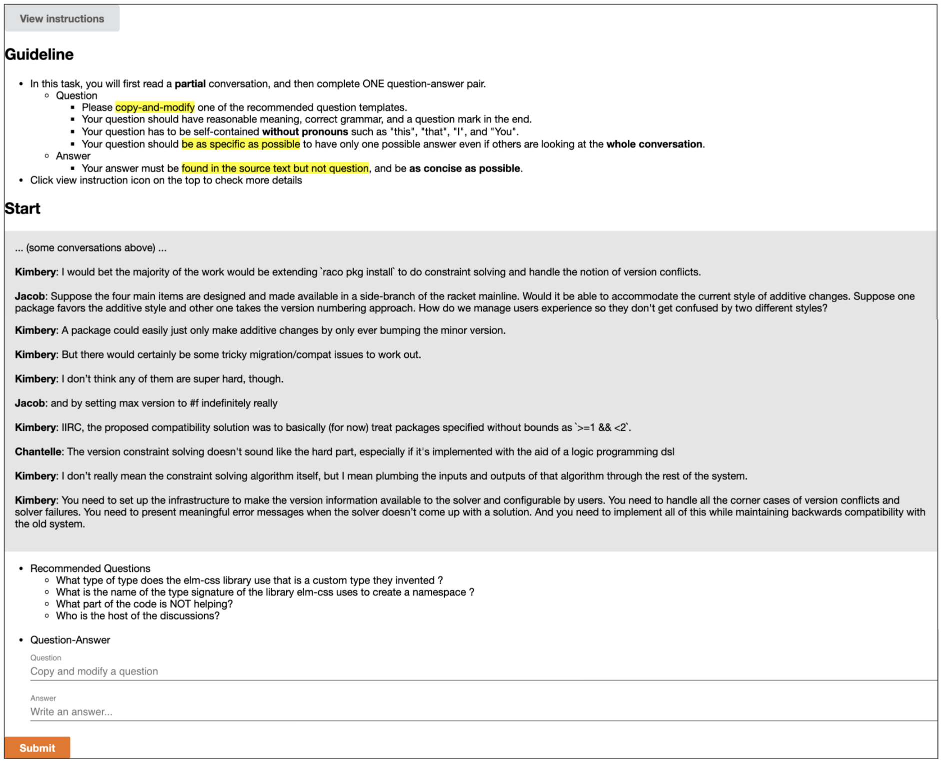
Terry Allen: My problem over the past year or so is that there are only a few session I wish to attend, but I cannot know for sure when they will be scheduled, so I cannot make reasonable travel arrangements (a week in Orlando for 6 hours of meetings is hard to sell to management). Now I know there is a rationale here, and that one is encouraged to participate broadly. And I am hopeful that new activities (my own and in the IETF) will give me many more reasons to attend. But firmer scheduling would be a big win.

- Question-Answer 1

Question
Write a question...

Answer
Write an answer...

Figure 4: Screenshot for human-written QA collection.



Guideline

- In this task, you will first read a **partial** conversation, and then complete ONE question-answer pair.
  - Question
    - Please copy-and-modify one of the recommended question templates.
    - Your question should have reasonable meaning, correct grammar, and a question mark in the end.
    - Your question has to be self-contained **without pronouns** such as "this", "that", "I", and "You".
    - Your question should be as specific as possible to have only one possible answer even if others are looking at the **whole conversation**.
  - Answer
    - Your answer must be found in the source text but not question, and be **as concise as possible**.
- Click view instruction icon on the top to check more details

Start

... (some conversations above) ...

Kimbery: I would bet the majority of the work would be extending `raco pkg install` to do constraint solving and handle the notion of version conflicts.

Jacob: Suppose the four main items are designed and made available in a side-branch of the racket mainline. Would it be able to accommodate the current style of additive changes. Suppose one package favors the additive style and other one takes the version numbering approach. How do we manage users experience so they don't get confused by two different styles?

Kimbery: A package could easily just only make additive changes by only ever bumping the minor version.

Kimbery: But there would certainly be some tricky migration/compat issues to work out.

Kimbery: I don't think any of them are super hard, though.

Jacob: and by setting max version to #f indefinitely really

Kimbery: IIRC, the proposed compatibility solution was to basically (for now) treat packages specified without bounds as `>=1 && <2`.

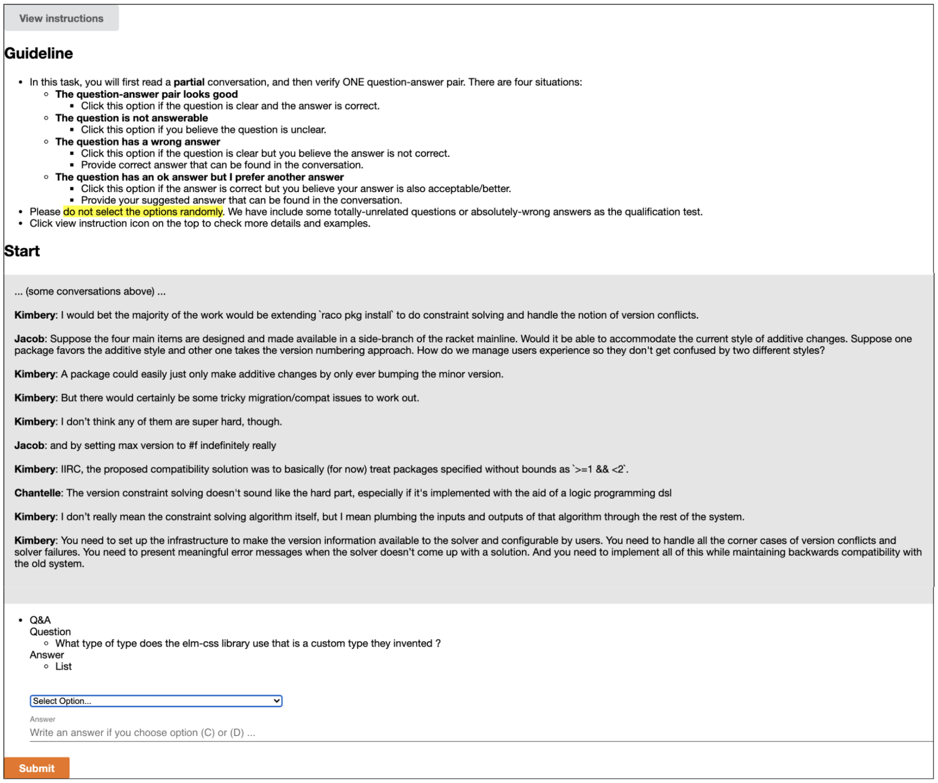
Chantelle: The version constraint solving doesn't sound like the hard part, especially if it's implemented with the aid of a logic programming dsl

Kimbery: I don't really mean the constraint solving algorithm itself, but I mean plumbing the inputs and outputs of that algorithm through the rest of the system.

Kimbery: You need to set up the infrastructure to make the version information available to the solver and configurable by users. You need to handle all the corner cases of version conflicts and solver failures. You need to present meaningful error messages when the solver doesn't come up with a solution. And you need to implement all of this while maintaining backwards compatibility with the old system.

- Recommended Questions
  - What type of type does the elm-css library use that is a custom type they invented ?
  - What is the name of the type signature of the library elm-css uses to create a namespace ?
  - What part of the code is NOT helping?
  - Who is the host of the discussions?
- Question-Answer

Figure 5: Screenshot for machine-generated QA collection.



Guideline

- In this task, you will first read a **partial** conversation, and then verify ONE question-answer pair. There are four situations:
  - **The question-answer pair looks good**
    - Click this option if the question is clear and the answer is correct.
  - **The question is not answerable**
    - Click this option if you believe the question is unclear.
  - **The question has a wrong answer**
    - Click this option if the question is clear but you believe the answer is not correct.
    - Provide correct answer that can be found in the conversation.
  - **The question has an ok answer but I prefer another answer**
    - Click this option if the answer is correct but you believe your answer is also acceptable/better.
    - Provide your suggested answer that can be found in the conversation.
- Please do not select the options randomly. We have include some totally-unrelated questions or absolutely-wrong answers as the qualification test.
- Click view instruction icon on the top to check more details and examples.

Start

... (some conversations above) ...

Kimbery: I would bet the majority of the work would be extending `raco pkg install` to do constraint solving and handle the notion of version conflicts.

Jacob: Suppose the four main items are designed and made available in a side-branch of the racket mainline. Would it be able to accommodate the current style of additive changes. Suppose one package favors the additive style and other one takes the version numbering approach. How do we manage users experience so they don't get confused by two different styles?

Kimbery: A package could easily just only make additive changes by only ever bumping the minor version.

Kimbery: But there would certainly be some tricky migration/compat issues to work out.

Kimbery: I don't think any of them are super hard, though.

Jacob: and by setting max version to #f indefinitely really

Kimbery: IIRC, the proposed compatibility solution was to basically (for now) treat packages specified without bounds as `>=1 && <2`.

Chantelle: The version constraint solving doesn't sound like the hard part, especially if it's implemented with the aid of a logic programming dsl

Kimbery: I don't really mean the constraint solving algorithm itself, but I mean plumbing the inputs and outputs of that algorithm through the rest of the system.

Kimbery: You need to set up the infrastructure to make the version information available to the solver and configurable by users. You need to handle all the corner cases of version conflicts and solver failures. You need to present meaningful error messages when the solver doesn't come up with a solution. And you need to implement all of this while maintaining backwards compatibility with the old system.

- Q&A
  Question
  - What type of type does the elm-css library use that is a custom type they invented ?

  Answer
  - List

Figure 6: Screenshot for QA verification.

Table 15: GPT3 zero-shot format. We prepend one conversational QA example from CoQA to samples of QAConv test set. We found the results are significantly better than using the QAConv data as prompt.

Prompt1 (CoQA)	Helsinki is the capital and largest city of Finland. It is in the region of Uusimaa, in southern Finland, on the shore of the Gulf of Finland. Helsinki has a population of, an urban population of , and a metropolitan population of over 1.4 million, making it the most populous municipality and urban area in Finland. Helsinki is some north of Tallinn, Estonia, east of Stockholm, Sweden, and west of Saint Petersburg, Russia. Helsinki has close historical connections with these three cities. The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns. It is the world's northernmost metro area of over one million people, and the city is the northernmost capital of an EU member state. The Helsinki metropolitan area is the third largest metropolitan area in the Nordic countries after Stockholm and Copenhagen, and the City of Helsinki is the third largest after Stockholm and Oslo. Helsinki is Finland's major political, educational, financial, cultural, and research center as well as one of northern Europe's major cities. Approximately 75% of foreign companies that operate in Finland have settled in the Helsinki region. The nearby municipality of Vantaa is the location of Helsinki Airport, with frequent service to various destinations in Europe and Asia. Q: what is the most populous municipality in Finland? A: Helsinki Q: how many people live there? A: 1.4 million in the metropolitan area Q: what percent of the foreign companies that operate in Finland are in Helsinki? A: 75% Q: what towns are a part of the metropolitan area? A: Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns
Prompt2 (QAConv)	NEAL CONAN, HOST: Lucy Morgan, what do you think? LUCY MORGAN: I think Romney will pull it out. The newest poll, the better of the polls, Quinnipiac, came out this morning giving Romney a one-point advantage, within the margin of error. But I think the advantage he has is the early vote and the establishment Republicans who are behind him. NEAL CONAN, HOST: Let's see if we can get some callers on the line, 800-989-8255. Florida primary voters, have you made up your minds yet? We'll start with Marie(ph), and Marie's on with us from Fort Lauderdale. MARIE: Yes, good afternoon. NEAL CONAN, HOST: Afternoon, go ahead, please. MARIE: Yes, well, I have decided to shift gears. Instead of voting Democrat, I am voting for Mitt Romney because I am a very dissatisfied Democratic voter. As a Haitian-American, I have seen no change in the status of a lot of my friends as far as finances are concerned. A lot of them have lost their houses, and all the promises that were made during the campaign, none of them have been accomplished. So I am switching my gears to see if I can vote for capitalism. NEAL CONAN, HOST: And so vote for Mitt Romney. Why Mitt Romney, if you're going to switch to vote Republican, among the other Republicans? Q: Which contact number is available for callers on the line said by NEAL CONAN? A:
GPT3 (Prompt 2)	I'm sorry, I'm sorry. I'm sorry. I'm sorry. I'm sorry.
GPT3 (Prompt 1 + Prompt 2)	800-989-8255

Table 16: Examples of multi-span answers in QAConv

Relevant Context	Question	Answer
... David Klinger: There’s a term of art called awful, but lawful. So sometimes officers are involved in shootings that don’t really sound that good, but the law says it was an appropriate ...	what can be awful but lawful?	officer involved shootings
... one foreign government should not be able to come into our courts and enforce its sovereign power by using our courts to collect taxes from our citizens...	how do one foreign government should not be able to come into the courts and enforce its sovereign power?	by using the courts to collect taxes from the citizens.
... directly in your mutable set without worrying about it, since there can only be expansion in one module per visit to your module. so you’ll never end up with ‘’module‘ being returned for two different modules before your mutable set is emptied. gonzalo: so, to ...	how many expansions can be in one module per visit?	one expansion per visit

Table 17: QG v.s. HW questions: test set results

		Zero-Shot			Finetune		
		EM	F1	FZ-R	EM	F1	FZ-R
QG	DistilBERT-Base (SQuAD 2.0)	42.94	47.67	61.80	64.04	72.02	78.79
	BERT-Base (SQuAD 2.0)	34.90	40.17	56.51	65.18	73.50	79.57
	BERT-Large (SQuAD 2.0)	**54.79**	**60.46**	**70.99**	71.59	78.58	83.78
	RoBERTa-Base (SQuAD 2.0)	52.44	57.77	69.28	70.05	77.93	83.14
	RoBERTa-Large (SQuAD 2.0)	52.60	58.57	69.69	**73.05**	**80.97**	**85.69**
HW	DistilBERT-Base (SQuAD 2.0)	48.44	55.58	66.26	63.50	74.99	80.40
	BERT-Base (SQuAD 2.0)	46.98	54.81	65.67	67.02	77.81	82.94
	BERT-Large (SQuAD 2.0)	**64.46**	**72.26**	**78.51**	64.99	76.64	81.97
	RoBERTa-Base (SQuAD 2.0)	60.63	68.20	75.70	71.73	81.67	85.90
	RoBERTa-Large (SQuAD 2.0)	62.53	70.86	77.47	**75.47**	**85.11**	**88.74**

Table 18: Examples of multi-hop questions

Partial Context	... Steve Duffy: ..., but I don't know if Enron would even consider this. Studdert might have\nthe best feel for this. Separately, the defendant group will get back to us\non any offer they might be willing to make to settle just the Montana case,\nbut it appears that their real interest would be in a \"global\" deal. Any\ncomments? SWD Michael Burke: Steve, Stan and I have discussed this and we agree that Mike Moran should\ntake the lead and explore all aspects of an Enron Global deal. I know that\nyou will assist Mike in this endeavor. thanks, mike Steve Duffy: Sounds good. Mike Moran has the numbers for our Montana lawyers and I will\nassist him any way I can. The big question is whether Enron, as a whole,\nwould be willing to give up any protection they might still have under the\nold InterNorth policies. SWD ...
Question	What person has the numbers for the Montana lawyers and is best qualified to explore the deal?
Partial Context	... OFEIBEA QUIST-ARCTON, BYLINE: One woman we spoke to has lived here all her life. She was born here, married here, has children here. She said I'm going. I don't feel safe. You know, the ground was shaking when we heard those bombs. We don't feel ... JENNIFER LUDDEN, HOST: We are talking about the tensions and violence in Nigeria. We'll have more with NPR's Ofeibea Quist-Arcton from Nigeria, and also former Ambassador John Campbell coming up. We'll also talk with an activist from Nigeria. If you have questions, ... JENNIFER LUDDEN, HOST: This is TALK OF THE NATION from NPR News. I'm Jennifer Ludden. Nigeria has long faced challenges from corruption, an economy that relies on oil exports and simmering ethnic and religious tensions, tensions made evident in the recent series of bombings by Boko Haram, the militant ... JENNIFER LUDDEN, HOST: It's the latest crisis for President Goodluck Jonathan. We're talking today with Ofeibea Quist-Arcton, NPR's foreign correspondent, now in Kano, Nigeria; and John Campbell, former U.S. ambassador and political counselor to Nigeria. He's now a senior fellow for Africa policy studies at the Council on Foreign Relations. ...
Question	Who is the president of the country where Ofeibea quist-arcton is talking about the tensions and violence in Nigeria ?
Partial Context	... Karoline: are you using pytest? there are a couple of plugins for parallelization Valeri: Yes pytest Eliana: pytest-xdist is pretty good Valeri: What does that do? Karoline : yeah that and pytest-parallel are worth a look . basically they allow you to paralelize your tests Valeri: Okay Valeri: Will definitely look into those Valeri: Thanks <@Eliana><@Karoline>,taco, …
Question	What program allows the user to parallelize the tests and is recommended by Karoline?
Partial Context	… MR. FREEDMAN (RESPONDENT): … They both deserve the death penalty. They – they were – the prosecutors were aware that the – the death penalty is what stirs the pot here, and so they were urging somebody to be the shooter to get the death penalty. If this wasn't a death penalty case, I don't think they – it would have mattered who killed who. And so they were urging – JUSTICE KENNEDY: Well, I think there's quite a difference in – in case A where you say our position is that Stumpf was the shooter, pure and simple. That's it. In case B, they say we think Stumpf was the shooter. We're not 100 percent sure, but he should get the death penalty. The alternative is before the sentencer and the sentencer can make that determination. …
Question	Which person was mentioned as the shooter in case A and B?