Socratic Question Generation: A Novel Dataset, Models, and Evaluation

Beng Heng Ang,[1] Sujatha Das Gollapalli,[2] See-Kiong Ng[2]
[1] Integrative Sciences and Engineering Programme, National University of Singapore
[2] Institute of Data Science, National University of Singapore
bengheng.ang@u.nus.edu,{idssdg,seekiong}@nus.edu.sg

Abstract

Socratic questioning is a form of reflective inquiry often employed in education to encourage critical thinking in students, or to elicit awareness of beliefs and perspectives in a subject during therapeutic counseling. Specific types of Socratic questions are employed for enabling reasoning and alternate views against the context of individual personal opinions on a topic. Socratic contexts are different from traditional question generation contexts where "answer-seeking" questions are generated against a given formal passage on a topic, narrative stories or conversations.

We present *SocratiQ*, the first large dataset of 110K (question, context) pairs for enabling studies on Socratic Question Generation (*SoQG*). We provide an in-depth study on the various types of Socratic questions and present models for generating Socratic questions against a given context through prompt tuning. Our automated and human evaluation results demonstrate that our *SoQG* models can produce realistic, type-sensitive, human-like Socratic questions enabling potential applications in counseling and coaching.

1 Introduction

Researchers in Education and Psychology have recognized the role of cognitive biases in shaping a person's perspective towards learning and understanding (Azzopardi, 2021; Bautista, 2014; Tversky and Kahneman, 1974; Vittorio et al., 2021). Indeed, both pedagogical and counseling environments involve recognizing and alleviating any flawed cognitive biases in students/subjects through appropriate interventions by trained professionals (Bhardwaj et al., 2018). Socratic Questioning is one such intervention technique pervasive in Education and Psychotherapy (Bautista, 2014; Chew et al., 2019; Vittorio et al., 2022).

Socratic Questioning involves the use of specific types of probing questions that guide people into eliciting biases underlying their understanding of a topic in order to potentially enable alternative perspectives and further thoughts (Paul and Binker, 1990; Paul and Elder, 2019). Figure 1 illustrates example Socratic questions on a passage expressing an individual's views on "eating meat".

Thought/Passage: I believe that eating meat is ethically wrong. Since we can easily substitute meat with vegan food without much nutritional complications, we have no logical reasons to continue eating living animals. We should stop killing these animals.

Possible Socratic Questions:

1. But what about eating animals that pass on from natural causes?
2. Where is your source that says that vegan food and meat have the same nutrition value?
3. What will happen if we continue killing animals for our consumption?

Figure 1: Example Socratic Questions

In this paper, we study automatic Socratic Question Generation (*SoQG*) as a novel, multidisciplinary application area for question generation (QG) research in NLP. Consider for instance, the publicly-accessible datasets available for learning question and dialog generation models (Rajpurkar et al., 2016; Ramnath et al., 2021; Talmor et al., 2017; Trischler et al., 2017; Yang et al., 2018). These existing QG datasets were designed for machine comprehension and comprise of news articles, Wikipedia, and other well-written formal passages and questions whose answers are potentially expressed in the passages. Similarly, current dialog generation datasets focus on capturing conversations related to completing specific tasks (such as restaurant booking, customer service) or learning open-domain conversational chatbots for chit-chat (Byrne et al., 2021; Cui et al., 2020; Danescu-Niculescu-Mizil and Lee, 2011; Zhou et al., 2018). As such, these existing datasets do not comprise of contexts necessary for learning *SoQG*.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 147–165
May 2-6, 2023

Socratic contexts differ from the scenarios captured in existing datasets in the following significant ways: 1. *Question Context.* The contexts or passages express individual thoughts and personal opinions on specific topics; 2. *Question Objective.* The questions do not seek a "correct answer" and aim to provoke introspection and reflection from the question recipient; and 3. *Question Type.* The questions adhere to specific Socratic types that challenge the completeness and accuracy of the thought expressed in the context in various ways. For example, the question-1 in Figure 1 seeks an *alternative perspective*, while question-2 probes for *evidence* regarding a claim expressed in the passage, and question-3 invokes further thought on *implications* of a specific action mentioned in the passage.

In this paper, we present a first study on Socratic Question Generation (*SoQG*) and make the following contributions:

1. We describe *SocratiQ*,[1] a large dataset of ∼110k Socratic questions, question type labels, and their contexts to enable learning of *SoQG* models. We discuss the curation of *SocratiQ* from the large corpus of posts and replies available from the community discussion website Reddit.[2] Socratic question-type labels collected through crowdsourcing are included for a subset of questions in *SocratiQ*.

2. We study question-type prediction for *SocratiQ* using state-of-the-art deep learning models. For a given (question, context) pair, our best fine-tuned BERT classifier is able to identify its question type with a macro F1-score of 0.905. We use this highly-accurate classifier to provide question-type labels for all instances in *SocratiQ* for learning type-sensitive question generation.

3. We learn type-sensitive question generation based on Socratic question types. We extend state-of-the-art QG models based on GPT-2, T5, and ProphetNet to incorporate Socratic question types through *prompt-tuning*. Our models effectively generate realistic, relevant human-like Socratic questions as shown in automatic as well as human evaluation studies.

Through our findings as well as released resources, we hope to enable future research on QG and chatbot applications based on the Socratic Questioning paradigm in areas such as coaching and counseling.

Organization: We describe the details of creating *SocratiQ* in Section 2. Section 3 provides details of the models and baselines used to study *SoQG*. Our experimental setup, evaluations, and results are summarized in Section 4 after which we provide a brief summary of existing datasets and methods for QG in Section 5. Finally, we conclude the paper in Section 6 and present some limitations that can be addressed in future in Section 7.

2 Dataset Collection

To construct a dataset of contentious viewpoints and the questions challenging these viewpoints, we consider the social news, content, and discussions website Reddit, in particular, the subreddit *r/changemyview*, or CMV. The CMV platform is an active, targetted community described as "a place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue" with people specially encouraged to "enter with a mindset for conversation, not debate". Upon closer examination of the posts, we found that CMV posts pertain to various controversial topics (e.g. politics, media, culture, etc.) and, due to its very purpose of "change my view", subsequent comments often include questions aimed to evoke introspection and reflection (further discussed in Section 2.1).

We obtained the raw data from CMV using the "Pushshift Reddit" API provided by Baumgartner et al (2020). After removing moderator comments that are not relevant to the topic,[3] we identify questions from comments following a given post using regular expressions comprising of question cues (e.g. who/what/where/when/why/how), and lexical indicators such as '?'. Next, each identified question sequence is paired with the most relevant paragraph sequence (context) in the previous posts based on the similarity of the two sequences. Sentence BERT encodings (Reimers and Gurevych, 2019) were used for similarity computation. After manually inspecting several examples and their similarity values, we retain pairs having a similarity value above the threshold of 0.55. In total, we produce a dataset of $110,050$ English[4] (question, context) pairs from the CMV content generated during January 2013 and December 2021. On average, the number of sentences in each context is $4(\pm 2)$

[1] https://github.com/NUS-IDS/eacl23_soqg
[2] https://www.reddit.com/r/changemyview/
[3] e.g., "All comments that earned delta ... are listed here"
[4] https://tinyurl.com/yzkxb2mz

Question type	Description	Exemplar
Clarification	Question probing the ambiguities of a thought.	What do mean by ...?
Probing assumptions	Question probing the assumptions behind a thought.	Why do you assume ...?
Probing reasons and evidences	Question probing the justifications or concrete evidences that could have supported a thought.	How did you know that ...?
Probing implications and consequences	Question probing the impacts or implications of a thought.	If ..., what is likely to happen as a result?
Probing alternative viewpoints and perspectives	Question probing other possible viewpoints.	What else should we consider about ...?
Others*	Question unrelated to the question types above (e.g. rhetorical, irrelevant, and/or illogical questions, etc.)	Who wouldn't want to be rich?

Table 1: Description and exemplar for each Socratic Question-Type from Paul and Elder (1990; 2019). *We add the catch-all type Others to refer to questions that do not conform to Socratic categories.

comprising of $83(\pm53)$ words whereas questions are a sentence long with $12(\pm6)$ words. We refer to this collection of (question, context) pairs as *SocratiQ*.

2.1 Question Annotation

Practitioners of Socratic Questioning employ various types of questions to engage in discussions with their subjects. For example, a counselor may ask for further clarification of a given viewpoint, or request evidence based on which such a stance was reached. Similar to previous studies (Dinkins and Cangelosi, 2019; Neenan, 2009; Wilberding, 2021), we used the taxonomy of Paul and Elder (1990; 2019) for characterizing the different questions. Table 1 lists the types of questions from this taxonomy with their descriptions and exemplars.

To curate the question type information for *SoQG* (Section 3), we adopt a semi-automatic process by first collecting manual annotations on a representative sample of data and training an accurate classifier for annotating the full dataset.

Crowd Labeling Process: To reduce the manual annotation efforts and cover the range of different question types uniformly in *SocratiQ*, we employ the process suggested in previous works (Abdul-Mageed and Ungar, 2017): First, lexical cues from available exemplars for each type (See Table 1) are used to design regular expressions[5] to assign a tentative Socratic type to each question in the dataset. We then use these tentative label assignments to sample questions from all types uniformly. In this manner, we sampled a set of 600 (question, context) pairs for each type, resulting in a balanced subset of $3,600$ questions for which we obtain human-assigned Socratic question-type labels.

Our annotation task was set up on the crowdsourcing platform Amazon Mechanical Turk (Crowston, 2012),[6] following previous works on QA/QG dataset collection (Ko et al., 2020; Rajpurkar et al., 2016; Yang et al., 2018). Each (question, context) pair was examined by three independent crowdworkers who chose one unique question type from among the six types listed in Table 1.[7]

We ensure the *ethics, quality, and reliability* considerations for all our collected datasets as follows. On the AMT platform, we required the crowdworkers to have greater than 95% HIT approval rate, a minimum of 10,000 HITs, be located in the United States and score at least 80% on a qualification test we set up to be able to work on our task. Each worker was paid up to 0.25 USD per HIT based on the task. In total, we used the AMT platform to collect data for the classification task as well as two evaluation studies which are further described in Section 4. The *anonymity* and *privacy* of the crowdworkers was already ensured on the AMT platform. We chose the pay-per-HIT based on similar on-going tasks on AMT. The settings for the HIT approval rates, and location of the worker ensures the English language skills of the annotator and thereby, the *quality* of the dataset.

A total of 40 workers helped in creating our classification dataset. About 40% of the workers who attempted our qualification test gained the eligibility to work on our task. Through this qualification test, we ensure that their worker annotations are *reliable*. For the $3,600$ (question, context) pairs deployed on AMT, we were able to obtain majority

[5]Examples are included in Appendix B.

[6]https://www.mturk.com/

[7]AMT interface details are provided in Appendix E.

annotations ($\geq$ 2 agreement on labels) for about 3,169 pairs. The Fleiss Kappa (Fleiss, 1971) inter-annotator agreement value was $\kappa = 0.725$ indicating substantial agreement among the crowd annotators. Overall, about 1% of the annotated questions were marked with the "Others" class which includes meaningless or irrelevant questions, and questions that do not belong to the Socratic types. This low percentage highlights the effectiveness of our pre-processing pipeline in extracting relevant Socratic (question, context) pairs.

Annotating the full dataset: We obtained question types for the entire *SocratiQ* dataset using an accurate classifier trained with the AMT labeled dataset. We further detail our classification experiments in Section 4. To summarize, in both automatic and manual evaluation studies, the prediction F1 and accuracy of our best classifier are around 0.9. Predicted Socratic classes are used for training prompt-based QG models described in Section 3.

2.2 Dataset Analysis

The question type distribution of the manually-annotated subset of *SocratiQ* is shown in Figure 2. Here, we observe that questions probing for *implications and consequences* and *reasons and evidence* comprise the most-asked questions. Previous studies have highlighted that "people routinely ask clarifying questions" to make sure "they can better offer assistance to the original poster" (Rao and Daumé III, 2018). This aspect is also seen in our dataset where *clarification* questions form the next most-asked questions. This is followed by questions probing for *assumptions* and *alternative viewpoints* as these are more cognitively-challenging than clarification questions (Krathwohl, 2002) while the "Other" questions form a negligible fraction (1.3%) of the overall distribution. Based on corpus analysis with Latent Dirichlet Analysis (Blei et al., 2003),[8] some representative topics in *SocratiQ* include *Abortion, Politics, Taxes, Crimes, Veganism, Racism, and Religion*.

3 Socratic Question Generation

To establish baseline performance on *SocratiQ*, we follow the current practice in NLP by fine-tuning state-of-the-art large pretrained language models (PLMs) with *SocratiQ* for **So**cratic **Q**uestion **G**eneration (*SoQG*) (Brown et al., 2020; Peters et al., 2019; Yu et al., 2021). We use commonly employed transformer-based language models, namely, Generative Pre-trained Transformer or GPT (Radford et al., 2019), the text-to-text transfer transformer or T5 (Raffel et al., 2019)), and ProphetNet (Qi et al., 2020). All three models yield relatively high performance on various NLP tasks (Wolf et al., 2020) including QG on other datasets (Chan and Fan, 2019; Ko et al., 2020).

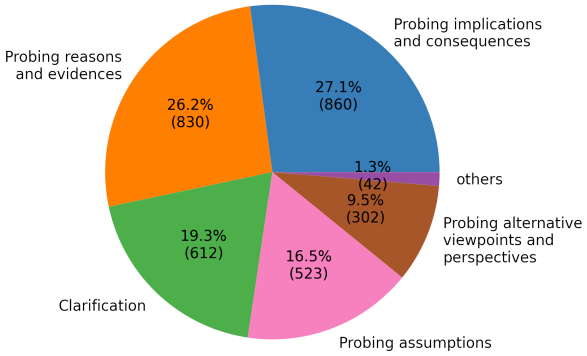

Figure 2: Distribution of questions based on their Socratic types

QG models: Following standard answer-agnostic QG models (Scialom et al., 2019), for our first set of baselines (GPT, T5, ProphetNet), we directly fine-tuned the PLMs with *SocratiQ* paragraph contexts as inputs and questions as outputs. We devised a second set of models (GPT-p, T5-p, ProphetNet-p) by employing prompt-based learning in keeping with recent developments in controllable text generation (Carlsson et al., 2022; Lester et al., 2021; Zhang et al., 2022). In this second set of models, we use the Socratic question-types as prompts in addition to the paragraph contexts as inputs for learning QG (Figure 3).

Evaluation: We use standard n-gram based metrics used for measuring question generation performance (Pan et al., 2019), namely BLEU (Papineni et al., 2002a), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004a). Previous works have highlighted the problems of using n-gram based measures for QG (Nema and Khapra, 2018). For *SoQG* in particular, these measures are limited in handling equally valid paraphrases of the available reference questions. To address these limitations, we adopt recently-designed learnt metrics, namely, BERT_Score and BLEURT (Sellam et al., 2020; Zhang et al., 2020; Yuan et al., 2021). Metrics based on the BERT-based models (Devlin et al., 2019) were shown to provide robust evaluation for

[8]Topic analysis results are included in Appendix A

GPT Input	<tag> [reasons_evidence] <tag> The government are people ... kind of things. <delim> If people are fantastic at managing all kinds of things, why are they so bad at them when organized as a government?
T5 Input	reasons_evidence: The government are people ... kind of things.
T5 Target	If people are fantastic ... bad at them when organized as a government?
ProphetNet Input	<tag> reasons_evidence <tag> The government are people ... kind of things.
ProphetNet Target	If people are fantastic ... bad at them when organized as a government?

Figure 3: Training input for each QG model. Text in red shows is the prompt based on question type annotation. <delim> is the delimiter token for separating input from the output whereas <tag> highlights the prompt tokens.

text generation tasks such as summarization and translation.

Configuration	R	P	F1
RE	0.681	0.706	0.617
L_{train}	0.797	0.929	0.828
(RE_D) + L_{train}	0.779	0.779	0.778
$(L_{train} + U_{RE})$	0.708	0.691	0.627
SelfTrain	0.826	0.883	0.846
GANBERT	0.784	**0.930**	0.802
$L_{train} + U_{V*}$	**0.933**	0.887	**0.905**

Table 2: Question-Type Classification Performance

4 Experiments

Setup: We set up our question-type classification and question generation experiments on a single GPU NVIDIA Tesla V100 machine. For classification, all models are based on the BERT classifier (Devlin et al., 2019).[9] For QG experiments, we use the implementations provided by the transformers library for GPT-2[10] and T5[11] and ProphetNet.[12] On our experimental machine, all models take between $12 - 19$ hours for training depending on the task, specific model and the dataset used.[13]

4.1 Classification Results

We studied several configurations based on BERT for training our question-type classifier. Table 2 shows the performance of these configurations on 20% (randomly-selected) test split of the AMT-annotated dataset using macro-averages of standard measures Recall/Precision/F1 employed for multiclass classification. We used the larger (80%) portion of the dataset for training and validation. The *SocratiQ* dataset contains $D \approx 110$K (question, context) pairs from which about $L \approx 3.1$K instances were collectively annotated by crowdworkers (Section 2.1). Thus, the unlabeled data ($U = D \setminus L$) can also be used to improve the classification performance via semi-supervised learning methods.

In Table 2, we show the performance using regular expression patterns in the first row (RE row) and the performance of BERT in the basic setting (fine-tuning on L_{train}) in the second row. The "(RE_D) + L_{train}" row refers to fine-tuning BERT on the tentative labels obtained with regular expressions before fine-tuning on L_{train}, whereas in the fourth row ($L_{train} + U_{RE}$), we add the unlabeled data using predictions with regular expressions to the training dataset.

The next three rows show semi-supervised configurations that involve the use of the unlabeled data (U) during model training. The "SelfTrain" row shows the performance in the self-training mode where the predictions from basic BERT configuration (second row) was used to predict labels for U and then re-trained on the combined dataset (Du et al., 2021; Mukherjee and Awadallah, 2020). The performance with the recently proposed semi-supervised model for BERT, namely, GANBERT (Croce et al., 2020) is shown in the next row.[14] For the final configuration in the last row ($L_{train} + U_{V*}$), we used voting to choose the dominant label from predictions on U with all configurations and retrained our classifier (Bishop, 2006). We did not include the regular expression-based models (RE and $L_{train} + U_{RE}$) in voting due to their substantially lower performance.

All classification models were fine-tuned using both the context and question as inputs. It is worth noting that the F1 performance of using labels based on regular expressions is significantly lower than the basic BERT configuration. Using only regular expressions or adding labels from regular expressions to unlabeled data results in significantly lower test performances (first and fourth rows) highlighting the inadequacy of only relying on exemplar

[9] https://huggingface.co/bert-large-cased
[10] https://huggingface.co/gpt2-large
[11] https://huggingface.co/t5-large
[12] https://github.com/microsoft/ProphetNet
[13] The details of deep learning experiment configurations are included in Appendix D
[14] https://github.com/crux82/ganbert

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	BERT_Score	BLEURT
GPT-p	0.165	0.013	0.167	0.187	0.615	0.423
GPT	0.150	0.007	0.149	0.167	0.601	0.412
T5-p	0.172	0.017	0.170	**0.211**	**0.632**	**0.426**
T5	0.144	0.011	0.142	0.179	0.615	0.413
ProphetNet-p	**0.178**	**0.018**	**0.177**	0.208	**0.632**	0.425
ProphetNet	0.152	0.011	0.147	0.178	0.616	0.416

Table 3: QG performance is shown using standard measures and BLEURT and BERT_Score values. All models that use question-type prompts perform significantly better compared to their respective baselines (t(10k), p <.05).

and templates for identifying Socratic Question type information. The best-performing configuration is obtained through combining all models to obtain the dominant label for unlabeled examples, resulting in a significant jump in F1 values as shown in the last row, even though semi-supervised learning using self-training and GANBERT have individually yielded small and no improvements, respectively.

We also experimented with only question tokens as input to the models. In general, the performance using question only as input is moderately high with the macro-F1 decrease of 1-8% compared to when both the context and question are used. Indeed, when the scores of tokens from the context versus the question are computed for their attribution towards prediction using the Integrated Gradients method from Sundararajan et al. (2017), we found that the average attribution score of question tokens, 0.118, is significantly much larger than that of the context tokens, 0.033 (Cohen's D value (2013) of 1.58). This suggests that the input tokens from question have more discriminatory power for predicting question type. Indeed, using context-tokens only as input, our basic BERT configuration obtained a macro-F1 score of only 0.185 indicating that the judgement of question-type is, not surprisingly, highly dependent on the question tokens as was also indicated in previous studies on question-type identification (Li and Roth, 2002; Svikhnushina et al., 2022).[15]

4.2 Question Generation Results

Next, we evaluate the models described in Section 3 on *SocratiQ*. For the models using question-type information (GPT-p, T5-p, ProphetNet-p), we append the predictions from our best classifier (last row in Table 2) with the context and appropriate separator tokens. For the (GPT, T5, ProphetNet) baselines, the context alone forms the input. We randomly split 105K pairs of *SocratiQ* that have a Socratic question label into Train/Dev/Test portions in the ratio "80/10/10" and show the test performance of QG models in Table 3.

Automatic Evaluation: As the reference questions from Redditors are available in *SocratiQ*, we directly employ the n-gram measures for evaluating QG performance. The BLEU-1/BLEU-4 values are shown along with METEOR and ROUGE scores in Table 3. We note that by incorporating question types, the prompt-based models do significantly better than their non-prompt counterparts for all three PLMs: GPT, T5, and ProphetNet. However, the overall n-gram overlap with reference questions is very low with the best BLEU-4 score of 0.018 and the best ROUGE score value of 0.211. For comparison, the ProphetNet model could yield a BLEU-4 score ranging from $0.23 - 0.25$ on test splits of SQuAD dataset (a well-used dataset in QG research) (Rajpurkar et al., 2016; Qi et al., 2020). Unlike these factual questions, Socratic questions can be fairly diverse for a given context and question type. Rather than measuring n-gram overlap, we posit that measuring semantic similarity between a given reference and generated question using recent text generation metrics such as BERT_Score and BLEURT, would be more indicative of the models' performance.[16]

Indeed, as shown in Table 3, the values of these learnt metrics are significantly higher than the n-gram based measures, with ProphetNet-p being the best performing model. When the generated questions are manually examined, we found that the fine-tuned PLMs indeed generate human-like questions which are relevant to the given contexts even though they do not match the given reference questions. This aspect was confirmed in our manual evaluation studies described next.

[15] Further details are included in Appendix C

[16] F1 metric was used for BERT_score. The precise models in both cases are described in Appendix D

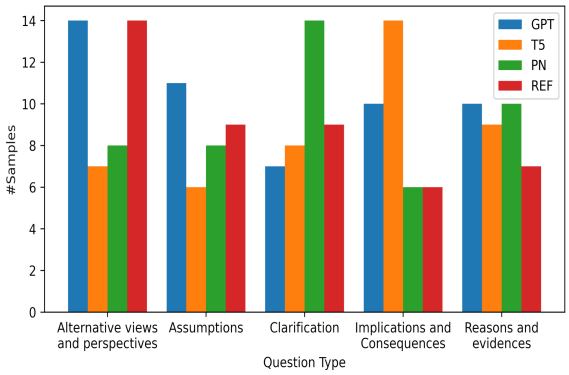

Figure 4: Distribution of "most likely to ask" questions chosen by crowdworkers

4.3 Human Evaluation Studies

We set up manual evaluation studies on Amazon Mechanical Turk to provide a more comprehensive analysis of *SoQG* model outputs. In the first study, we randomly sample fifty contexts for each Socratic question type of *SocratiQ*. For each context and Socratic question type, three independent MTurk evaluators select among four randomly-ordered questions (3 generated by GPT-p, T5-p, and ProphetNet-p + 1 reference) the question they will likely ask, or the "None" option.[17] In Figure 4, we plot the distribution of the source of the selected questions for each question type using majority votes from the above study.

Model	Flu	Rel	Ans
Best	4.287	3.883	0.330
Worst	4.220	3.737	0.353

Table 4: Quality of Questions (Flu:fluency, Rel:relevance, Ans:answerability)

We found that only two questions out of the 250 questions used in the study were deemed unacceptable for the given question-type and context by our annotators, confirming the accuracy of our question-type classifier. For the remaining questions, we see from Figure 4 that based on the question type, crowdworkers seem to prefer questions generated by a specific model and interestingly, reference questions from CMV were not always the clear winner. For example, the questions generated by ProphetNet were clearly preferred for the "Clarification" type of questions whereas T5-generated questions were preferred for "Probing implications and consequences" type. In fact, the human-reference questions were on par with the best model only in one type out of five indicating that our models are generating fluent and very human-like Socratic questions.

In a second related study, we sample twenty contexts for each of the five question types in the previous study, select the questions from the best and worst models (according to Figure 4), and ask three independent AMT crowdworkers to rate on a Likert scale of 1-5 ("Very bad" to "Very good") how fluent and relevant the questions are for a given context. They were also asked "if the answer to the question is present in the passage" (0/1 indicating absence/presence of the answer). The results for this study are summarized in Table 4. We found that the questions from the best models for each type indeed have higher fluency and relevance compared to questions from the worst model and both scores are close to the "Good" range on the Likert scale. Moreover, the answerability scores less than 0.5 indicate that the questions are not answerable from the context (different from traditional QG datasets).

Finally, we randomly sampled and examined fifty ground-truth questions per question type. These questions were examined on a 0/1 scale for classifier accuracy, fluency, answerability, and relevance. We found that the accuracy values ranged between 0.86-0.96, fluency and relevance between 0.92-1.00, and answerability was near zero (0.04), among the five Socratic classes. These scores further support our best classifier's performance on the AMT annotated test dataset, and that the ground-truth questions are fluent, relevant, and unanswerable from the context. These numbers indicate that our dataset of 110K (question, context) pairs and their predicted Socratic question types are of high quality and can be used by fellow NLP researchers for further improving *SoQG* models.

Anecdotes: The model-generated questions for an example context from *SocratiQ* are shown in Table 5 for illustration. We see that our prompt-based models are able to incorporate the question type information to generate questions that are semantically similar to the reference questions.

SQuAD versus *SocratiQ*: To highlight how Socratic questions complement existing QA/QG datasets, we run *SoQG* for an example from SQuAD (Rajpurkar et al., 2016).[18] For this context,

[17] Further details are included in Appendix E

[18] https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/Immune_system.html

Context: Beyond the boundaries of legal conduct, to exclude people from the society is judged to be a suitable place of the law. For some heinous conduct, it's acceptable to put people beyond the society, and exclude them for tye benefit of all. But the denial of voting rights to convicted felons after the end of thier sentence is not acceptable. **1 - Probing implications and consequences**: Are there serious risks to allowing felons to vote? **2 - Probing reasons and evidences**: Why are we concerned about felons voting and what value does purging felons from voter rolls have to our society?
GPT-*p* - 1: What will happen if the felons were convicted of a crime? **GPT-*p* - 2**: Why should felons be excluded from voting rights? **GPT**: Why should felons be denied the right to vote?
T5-*p* - 1: What will happen if the felons are not allowed to vote? **T5-*p* - 2**: Why is it acceptable to deprive people of their right to vote? **T5**: Is it acceptable to deprive convicted felons of their right to vote?
ProphetNet-*p* - 1: Are you implying that felons should be allowed to vote? **ProphetNet-*p* - 2**: Why is it acceptable to deny felons voting rights after the end of their sentence? **ProphetNet**: What about people who have been convicted of murder?

Table 5: Questions generated by the different models are shown for an example context and associated reference questions from *SocratiQ*. Table 12 of the Appendix contains some more examples.

note the contrast between the "fact-seeking" questions from SQuAD and Socratic questions listed in Table 6. Socratic questions focus on "what if/why/on what basis" angles, causing the reader to reflect on alternate perspectives and implications, which can be useful for nurturing further thought and analysis which are essential in Education.

Reference questions from SQuAD 1. What are the anitmicrobial peptides secreted by the skin called? 2. What enzymes in saliva are antibacterial in nature? 3. What compounds in the stomach protect against ingested pathogens? 4. Vaginal secretions serve as a chemical protective barrier following what?
T5-*p* generated Socratic Questions 1. How do these enzymes kill pathogen? 2. Is it a good thing to assume that the stomach and intestinal tract are chemically different? 3. Do you have evidence that these enzymes are effective at killing pathogen? 4. Are you implying that the stomach is more acidic than other organs?

Table 6: Example from the SQuAD Dataset

We present in Table 7 an example context from the Real-world Worry Dataset (Kleinberg et al., 2020). Against this context, *SoQG*-models generate questions that draw attention to the underlying assumptions and potential misconceptions related to the mentioned "worry" causing the reader to introspect, which can be useful during counseling.

Context: I am concerned for my family's safety and I'm worried about the impact isolation will have on the mental health of my loved ones. **GPT-*p* generated Socratic Questions** 1. What does isolation have to do with your mental health or safety? 2. Why do you assume that isolation will lead to mental health issues? 3. Have you considered that isolation is a mental health issue that affects the entire family?

Table 7: Example from Real-world Worry Dataset

5 Related Works

Question generation has been a prominent subject of recent NLP research (Pan et al., 2019; Lu and Lu, 2021). Except some QG datasets which discuss inquisitive, probing questions (Ko et al., 2020), clarification questions (Rao and Daumé III, 2018), and empathetic questions in social dialogs (Svikhnushina et al., 2022), most models are trained on datasets such as SQuAD (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), QuAC (Choi et al., 2018), and CoQA (Reddy et al., 2019), WebQA (Talmor et al., 2017) that were originally created for question answering (Pan et al., 2019). Not only are these datasets targeted towards extractive or abstractive QA and therefore comprise mostly factual, answer-seeking questions, but their contexts also include non-personal contexts such as Wikipedia or news articles, narrative stories, or problem descriptions on forums such as Stack Exchange (Rao and Daumé III, 2018). In contrast, Socratic contexts involve personal and individual opinions and viewpoints on diverse topics with no fixation on "correct" answer (Paul and Elder, 2019).

For learning QG models, we adopt the current approach of fine-tuning transformers on specific datasets (Wolf et al., 2020; Kriangchaivech and Wangperawong, 2019). Though *SoQG* is an answer-agnostic QG task (Scialom et al., 2019), due to its uniqueness in availability of question type information, we extend our QG models to incorporate these cues according to latest research on conditional text generation through the use of prompts (Lester et al., 2021; Zhang et al., 2022; Carlsson et al., 2022). Question-type taxonomies were previously studied for factual questions (Li

and Roth, 2002) and in context of social dialog (Svikhnushina et al., 2022).

6 Conclusions

We created a novel dataset *SocratiQ* to support research on automatic Socratic Question Generation. We applied latest research in prompt-based conditional text generation to fine-tune existing large language models from GPT, T5, and ProphetNet to learn *SoQG*. Through our study and the release of this novel dataset, we take a first step towards enabling future research on models for *SoQG* as well as impactful applications in areas such as counseling and education (Inkster et al., 2018; Fitzpatrick et al., 2017).

7 Limitations

We note the following limitations in our work that also comprise our future research directions. First, while a human Socratic method practitioner will know what type of Socratic question to ask based only on context, our prompt-based models assume the availability of question-type for generating a type-sensitive question. In fact, when only contexts were used for QG (GPT, T5, ProphetNet baselines in Section 3), the generated questions matched the desired question-type (those of the available reference questions) in only 37-40% of the cases. Furthermore, the question-type identification of automated methods using context alone was very poor with overall accuracy comparable to that of random assignments (Section 4.1).

Secondly, though we showcased the potential use of *SoQG* in designing chatbots and dialog systems for applications such as counseling, we note that the current evaluation has only been at the single-turn level. We hope to extend *SocratiQ* to capture back and forth discussions on CMV to provide multiturn data and also deduce via forum votes and other indicators if the discussion indeed resulted in changed minds and enabled alternate perspectives. Furthermore, considering the special purpose of Socratic questions in shaping perspectives and enabling introspection and reflection, a comprehensive evaluation would require measuring these aspects over the multi-turn sessions.

Finally, our dataset was created by re-purposing the CMV subreddit data available in English, a high-resource language for which large-scale pretrained language models (PLMs) are readily available. Obtaining high classification and generation performances via fine-tuning of PLMs will be a challenge that needs addressing in low-resource languages.

Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund–Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We thank our reviewers for their suggestions on improving the presentation of this paper and insights on future possibilities.

Ethics Statement

This research was conducted in accordance with the ACM Code of Ethics. The ethical considerations during the dataset collection process are discussed in Section 2.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 27–37.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Lowell Bautista. 2014. The socratic method as a pedagogical method in legal education. *Faculty of Law, Humanities and the Arts - Papers. 1481.*

Gaurab Bhardwaj, Alia Crocker, Jonathan Sims, and Richard D Wang. 2018. Alleviating the plunging-in bias, elevating strategic problem-solving. *Academy of Management Learning & Education*, 17(3):279–301.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, and Mihir Kale. 2021. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 671–680, Online. Association for Computational Linguistics.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857, Dublin, Ireland. Association for Computational Linguistics.

Devrim Cavusoglu, Fatih Cagatay Akyon, Ulas Sert, and Cemil Cengiz. 2022. Jury: Comprehensive NLP Evaluation toolkit.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Sie Wai Chew, I-Hsiu Lin, and Nian-Shing Chen. 2019. Using socratic questioning strategy to enhance critical thinking skill of elementary school students. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 290–294. IEEE.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 1406–1416, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

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

Christine Sorrell Dinkins and Pamela R Cangelosi. 2019. Putting socrates back in socratic method: Theory-based debriefing in the nursing classroom. *Nursing Philosophy*, 20(2):e12240.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Loshchilov Ilya and Hutter Frank. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR), 2019*.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Intel. 2007. Designing effective projects: Questioning the socratic questioning technique.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555, Online. Association for Computational Linguistics.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Chin-Yew Lin. 2004a. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chao-Yi Lu and Sin-En Lu. 2021. A survey of approaches to automatic question generation:from 2019 to early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 151–162, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Subhabrata (Subho) Mukherjee and Ahmed H. Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *NeurIPS 2020 (Spotlight)*. ACM.

Michael Neenan. 2009. Using socratic questioning in coaching. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 27(4):249–264.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *CoRR*, abs/1905.08949.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Richard Paul and Linda Elder. 2019. *The thinker's guide to Socratic questioning*. Rowman & Littlefield.

Richard W Paul and AJA Binker. 1990. *Critical thinking: What every person needs to survive in a rapidly changing world.* ERIC.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

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

Kiran Ramnath, Leda Sari, Mark Hasegawa-Johnson, and Chang Yoo. 2021. Worldly wise (WoW) - cross-lingual knowledge fusion for fact-based visual spoken-question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1908–1919, Online. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2736–2745.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032, Florence, Italy. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973, Dublin, Ireland. Association for Computational Linguistics.

Alon Talmor, Mor Geva, and Jonathan Berant. 2017. Evaluating semantic parsing against a simple web-based question answering model. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 161–167, Vancouver, Canada. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Lisa N Vittorio, Justin D Braun, Jennifer S Cheavens, and Daniel R Strunk. 2021. Cognitive bias and medication use moderate the relation of socratic questioning and symptom change in cognitive behavioral therapy of depression. *Cognitive Therapy and Research*, 45(6):1235–1245.

Lisa N Vittorio, Samuel T Murphy, Justin D Braun, and Daniel R Strunk. 2022. Using socratic questioning to promote cognitive change and achieve depressive symptom reduction: Evidence of cognitive change as a mediator. *Behaviour research and therapy*, page 104035.

Erick Wilberding. 2021. *Socratic methods in the classroom: Encouraging critical thinking and problem solving through dialogue*. Routledge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *CoRR*, abs/2201.05337.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Supplementary Materials

A Composition of *SocratiQ*

We apply Latent Dirichlet Analysis using the topic modeling toolkit, Mallet (McCallum, 2002).[19] The top-10 topics (by their Topic Coherence values (Mimno et al., 2011)) uncovered in *SocratiQ* with LDA (number of topics set to 30) are shown with their top words in Table 8. Based on these words, we note that some representative topics in *SocratiQ* include *Abortion, Politics, Taxes, Gender, Crimes, Veganism, Racisim, and Religion*.

	Top Words
1	child abortion life human children fetus woman baby parents mother
2	vote trump party voting president election states political democrats republicans
3	money people pay work tax make taxes government income wage company
4	women men gender trans woman male sex man female people transgender
5	crime rape people death crimes punishment person prison murder law
6	animals food meat eat animal eating humans dog dogs killing
7	people white black racist racism race culture person racial group
8	im view people dont ive change argument opinion post point
9	religion god religious christian bible religions islam people beliefs christianity
10	school college students education schools student high class job degree

Table 8: Words from the Top-10 topics (as ranked by Topic Coherence) from LDA analysis are shown

Figure 5 illustrates the distribution of question types predicted by our best classifier over *SocratiQ*. Compared to the manually annotated spread shown in Figure 4, we note that the percentage of Clarification type questions is significantly higher (33% versus 19%). We attribute this to the fact that Clarification class seems to be the most-confused one among the others as shown in the confusion matrix on the test set presented in Figure 6. As such, even our best-performing classifier is not 100% accurate.

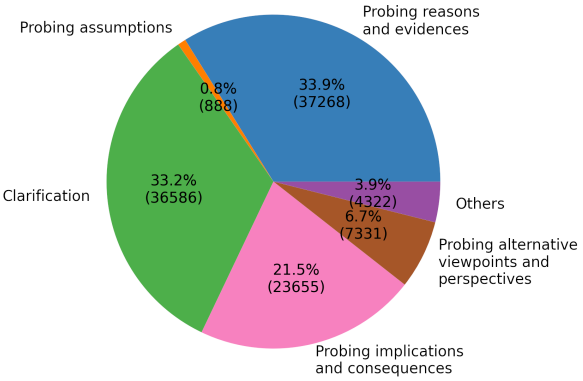

Figure 5: Distribution of predicted question types in the *SocratiQ*

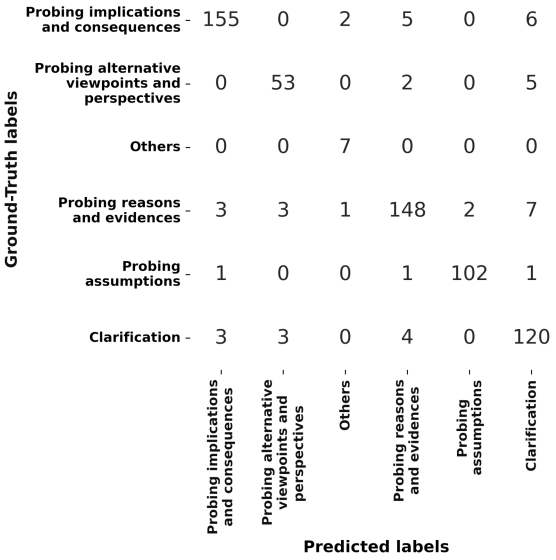

Figure 6: Confusion matrix of best classifier's predictions on the test set

[19] https://mallet.cs.umass.edu

Question type	Lexical Cues	Regex
Clarification	What do mean by ...?/What is the meaning of ...?	[Ww]hat[\s\w]+mean
	How is ... related to ...?/How are they related?	[Hh]ow[\s\w]+relate
Probing Assumptions	Are you assuming that...?/Why are you assuming that...?	[Aa]re[\s\w]+assum[eing]+
	What is the assumption from...	[Ww]hat[\s\w]+assumption
Probing reasons and evidences	Where is your evidence that...?	[Ww]here[\s\w]evidence
	Why do you think...?	[Ww]hy[\s\w]+think
Probing implications and consequences	Are you implying...?	[Aa]re[\s\w]+impl[yied]
	What happens...?What would have happened if...?	[Ww]hat[\s\w]+happen
Probing alternative viewpoints and perspectives	What else...?	[Ww]hat[\s\w]+else
	What other...?/What are the other...?	[Ww]hat[\s\w]]+other

Table 9: Sample lexical cues and regular expression patterns used to tentatively map questions to their Socratic question types.

B Regular Expressions

Table 9 provides sample lexical cues and their corresponding regular expressions (regex) designed based on available Socratic question templates (Paul and Elder, 2019; Intel, 2007). We use them to tentatively assign labels to instances in *SocratiQ*. These labels are used to get more balanced samples of questions for the human annotation process described in Section 2. The full set of regular expressions will be made available through our code repository.

C Computing Attributions for Classifier Predictions

We explain the model predictions by applying the Integrated Gradients method (Sundararajan et al., 2017).[20] For a question type prediction, each token is given an attribution score that indicates its contribution to that prediction. We normalize the scores by averaging them over the sequence length of each question and context sequences. From Table 10, we observe that the attribution values from the question sequences is significantly higher than those from the contexts; thereby, suggesting that the classifier predicts the question types mainly based on tokens from the question.

Based on the attribution scores, we can extract textual patterns that our classifier associates with particular question types in a data-driven fashion by simply using the top N-grams with the highest attribution scores. We show sample 4-grams in Table 11. We note the similarities between these automatically extracted patterns with the template-

Question Type	Qseq attr	Cseq attr	Cohen's D
Overall	.118	.033	1.59**
Clarification	.148	.041	1.85**
Probing reasons and evidences	.125	.037	1.64**
Probing implication and consequences	.108	.032	1.51**
Probing assumptions	.087	.020	1.76**
Probing alternate viewpoints and perspectives	.110	.025	1.80**
Others	.131	.061	0.76**

Table 10: Attribution scores (attr) for question sequences remain consistently higher than those of context sequences across every question type. Qseq indicates "Question Sequence" and "Cseq" indicates "Context Sequence". ** indicates a statistically significant, $p < 0.05$, Cohen's D effect size (Cohen, 2013) of the difference in attributions.

[20] https://captum.ai/

exemplars used by existing Socratic practitioners listed in Table 1.

GPT Input	<tag> [reasons_evidence] <tag> The government are people ... kind of things. <delim> If people are fantastic at managing all kinds of things, why are they so bad at them when organized as a government?
T5 Input	reasons_evidence: The government are people ... kind of things.
T5 Target	If people are fantastic ... bad at them when organized as a government?
ProphetNet Input	<tag> reasons_evidence <tag> The government are people ... kind of things.
ProphetNet Target	If people are fantastic ... bad at them when organized as a government?

Figure 7: Training input for each QG model. Text in red shows is the prompt based on question type annotation. <delim> is the delimiter token for separating input from the output whereas <tag> highlights the prompt tokens.

D Configuration Details for Deep Learning Experiments

For our classification experiments in Section. 4.1, we fine-tuned the "bert-large-cased"[9] model for the various configurations (Table 2). The model is tuned by an AdamW optimizer (Ilya and Frank, 2019) set with betas default at (0.9, 0.999), a batch of 4, and an initial learning rate of 1e-6. We also use a ReduceLROnPlateau[21] learning rate scheduler to reduce the learning rate by a default factor of 0.1 whenever the F1-score from the validation set does not improve after 2 epochs. Following Batista et al. (2004), we randomly oversample the minority classes to alleviate class imbalance among the data.

For our question generation experiments in Section 4.2, we fine-tuned "gpt-large",[10] "t5-large",[11] and "Prophetnet-En"[12] models. Prompt-tuning is enabled by concatenating question type annotations before the input sequence as shown in Figure 7. For "gpt-large" and "t5-large", we fine-tune the models using the pytorch [22] and huggingface[23] libraries, with an AdamW optimizer (Ilya and Frank, 2019) of betas (0.9, 0.999), a batch of 4, and an initial learning rate of 5e-5. This learning rate is adjusted during training with a ReduceLROnPlateau learning rate scheduler that reduces the learning rate by a factor of 0.1 whenever the loss from the validation set stops decreasing after 2 epochs. Similar to Ko et al (2020), for the "gpt-large" language model, we accumulate losses only for the question tokens by masking the context tokens before the delimiter. For ProphetNet, we use the recommended hyperparameters[12] except the learning rate, batch size,[24] input and target sequence lengths. For our models, these sequence lengths are set at 400 and 80 respectively to account for computation-related limits on our experimental machine.

For BLEU (Papineni et al., 2002b), ROGUE (Lin, 2004b), and METEOR (Banerjee and Lavie, 2005) automatic evaluation metrics, we use the implementations provided in Jury (Cavusoglu et al., 2022).[25] For BLEURT (Sellam et al., 2020), we use their recommended BLEURT-20 model.[26] Similarly, we use the recommended "microsoft/deberta-xlarge-mnli"[27] model for the BERTScore metrics.

E Details of MTurk Crowdsourced Tasks

Qualification Task: We presented a qualification test to crowdworkers on Amazon Mechanical Turk (AMT) who are interested in working on our tasks. The descriptions for each Socratic type along with examples are provided as instructions (Figure 8). As part of the test, the workers assign question type to a set of twelve questions (two for each type from Table 1). A snapshot from AMT platform for the qualification test is shown in Figure 9. We paid the workers, 0.05 per HIT for the classification task, 0.15 per HIT for the task involving question selection, 0.25 per HIT for the task involving question quality annotations on the Likert scale. These values were selected based on relevant, similar tasks on the AMT platform at the time of deployment.

The snapshots from AMT for the question selection and question quality studies described in Section 4 are shown in Figures 10 and 11.

[21] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

[22] https://pytorch.org/

[23] https://huggingface.co/

[24] Learning rate and batch size are set 5e-5 and 4 in consistent to "gpt-large" and "t5-large"

[25] https://pypi.org/project/jury/

[26] https://github.com/google-research/bleurt

[27] https://github.com/Tiiiger/bert_score

Clarification	Probing assumptions	Probing reasons and evidences	Probing implications and consequences	Probing alternative viewpoints and perspectives	Others
How does that definition... (0.467)	How can one conclude... (0.439)	Why did that matter... (0.461)	Should we not care... (0.441)	Or what about drinking... (0.441)	Did you not listen... (0.431)
What do they mean... (0.457)	Is your hypothesis that... (0.385)	Was there a reason... (0.412)	Would it cause people... (0.434)	How about other physical... (0.434)	Who even cares about... (0.416)
Do you mean doctors... (0.441)	Would you have assumed... (0.376)	Can you list any... (0.408)	Would the situation change... (0.429)	What other species would... (0.424)	Do you not realise... (0.391)
How does that equal... (0.430)	What makes you presume... (0.360)	What level of evidence... (0.357)	Are you implying rich... (0.426)	Was there anything else... (0.415)	Did you not understand... (0.370)
How does this agree... (0.413)	Why does society assume... (0.338)	What long term evidence... (0.354)	What happens when inclusion ... (0.413)	How would any other... (0.367)	What kind of dumbass... (0.347)

Table 11: 4-grams and their attributions (in brackets) for the different Socratic question types

CONTEXT:A child should hold no religious (or absense of religion) position. It is as unsettling as calling a child a conservative or liberal child. Children are vulnerable youths that take their parental figures' words as absolute fact. Telling them to hold political views or religious views at a young age is borderline indoctrination. **References**:1. What is the harm to the child by raising them as a liberal or a Muslim or a vegetarian or a meat eater? 2. What about teaching your child morals? **Generated Questions**: (Implications/Consequences) What do you think happens to young children who are forced to hold extremist views? (Clarification) How does teaching kids about the bible relate to holding religious views? (Reasons/Evidence) Why does it matter what political views children hold at such a young age? (Assumptions) Why are you assuming kids are even being taught their political views at all? (Alternate Viewpoints/Perspectives) Or that you would rather parents continue to influence them through their actions instead of just giving them the information?
CONTEXT:Adult children 25 or older should have to contribute and carry their weight. It would be on a case by case basis but essentially the child would have to pay a certain amount according to what they make. If you make over this much, you must pay 50% of the rent or mortgage. If you make below this much, you pay 15-25% and so on and so forth. This would encourage adults to grow up, pack up and move on from their parents homes. **References**: 1. What about parents who live at their kids home? 2. What will happen if the parents own the house free and clear? **Generated Questions**: (Implications/Consequences) Are you implying the bill could be identical between an adult and a child who are not of the same age ? (Clarification) What part of the point would you have that will encourage adults to move on? (Reasons/Evidence) And why should a person be required to move out at all? (Assumptions) Why do you assume that adults would just move on from their parents when they are adults? (Alternate Viewpoints/Perspectives) What about people who stay at home and take care of their parents and siblings for the rest of their lives?
CONTEXT:As a non-American, i find the idea of colleges having, promoting and sponsoring young (mostly female) students to dress in skimpy outfits to perform dangerous stunts and basically serve as eye-candy for the players and the audience a disturbing concept in the very least. Now don't get me wrong, the stunts they do, they teamwork they show are commendable and in no way a lesser sport in themselves. I'm sure many of the cheerleaders could have great careers in gymnastics or such organized sports. However, there is a drastic difference in how a cheerleading squad trains versus how someone trains for other sports. The cheerleaders often don't have safety equipment, practice on hard surfaces and have little to no health care. **References**:1. If the cheerleaders felt demeaned why would they choose to keep doing it, especially into college? 2. Have you consulted those cheerleaders? **Generated Questions**: (Implications/Consequences) Are you asking the college cheerleaders to do something about it? (Clarification) How do cheerleaders train on a surface that hard? (Reasons/Evidence) Do you have any evidence to support the idea that cheerleaders in colleges might be underprivileged? (Assumptions) Are we assuming cheerleaders are not working out to their maximum potential? (Alternate Viewpoints/Perspectives) What about professional gymnasts?

Table 12: Examples from *SocratiQ* are shown with their reference questions and questions generated with the ProphetNet$-p$ model

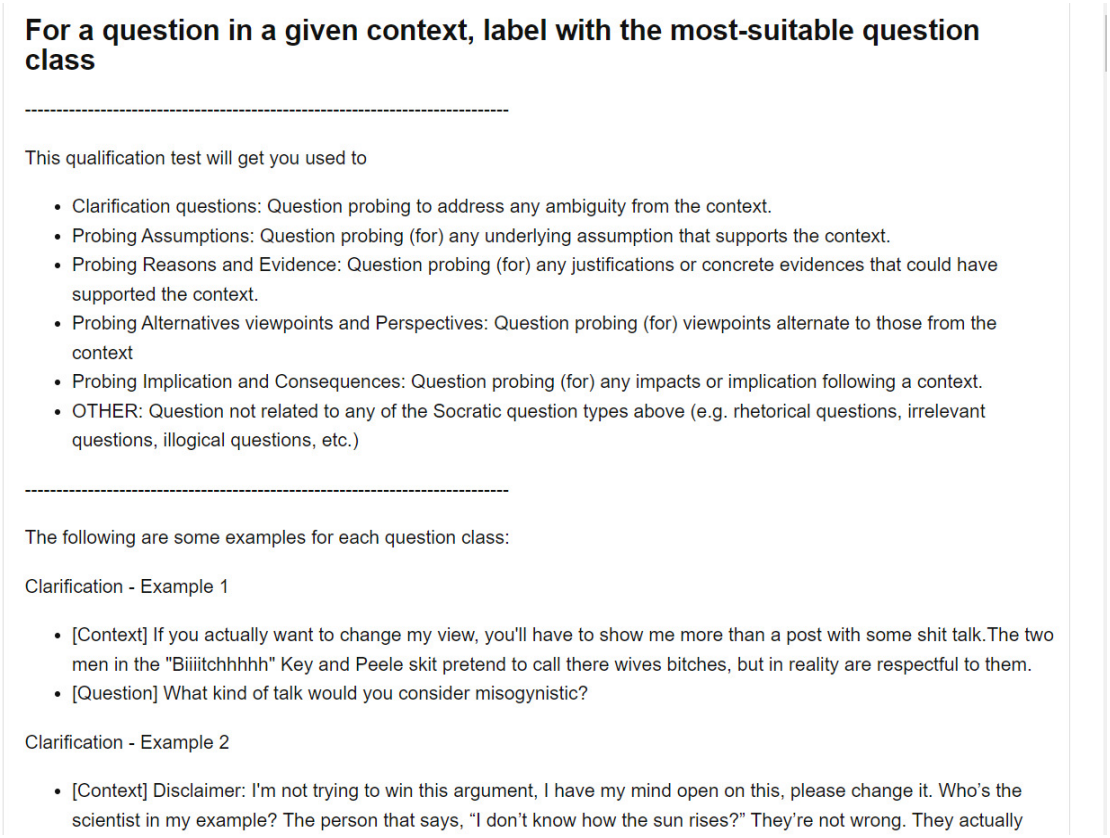

Figure 8: Workers are provided materials and time to learn the question types used in our study

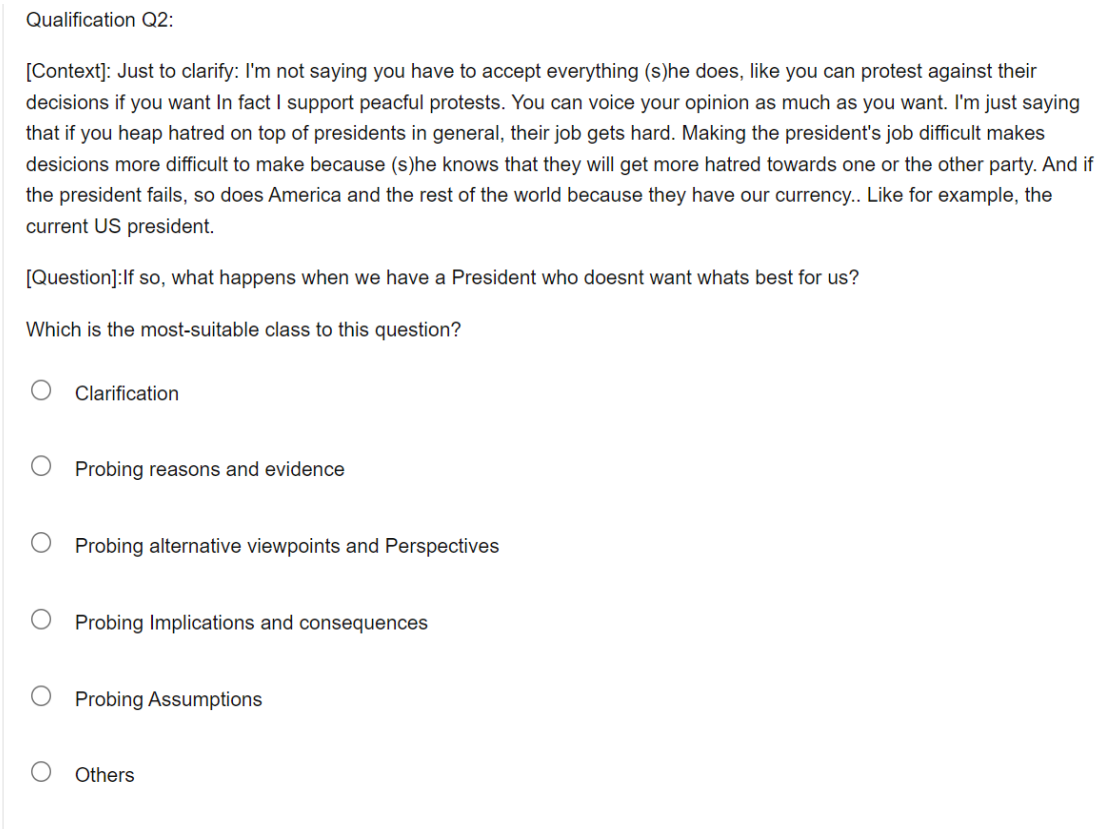

Figure 9: Workers choose a particular question type most suited to a (question-context) pair in the qualification test (as in the main task).

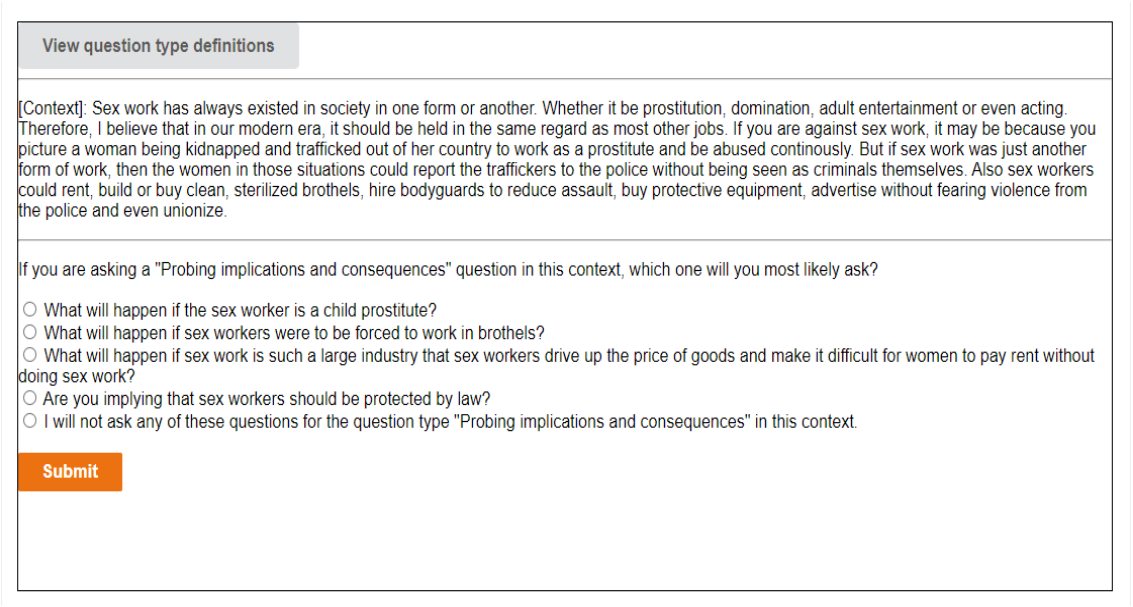

Figure 10: Interface on AMT for the question selection study

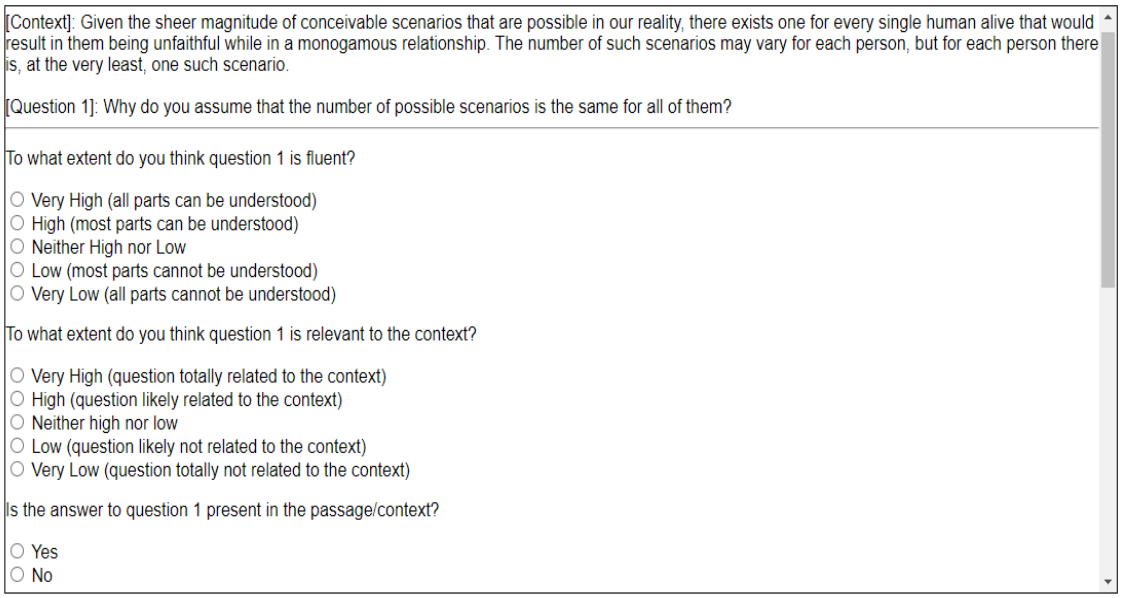

Figure 11: Interface on AMT for the question quality study