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 elicitings 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”.

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 chitchat (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.
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**, 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. **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, 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 (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(±2)
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.

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

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.

Our annotation task was set up on the crowdsourcing platform Amazon Mechanical Turk (Crowston, 2012), 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 assigned Socratic question-type labels.

---

\(^5\)Examples are included in Appendix B.
\(^6\)https://www.mturk.com/
\(^7\)AMT interface details are provided in Appendix E.
annotations (≥ 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), 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 Socratic Question Generation (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).

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 controlable 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

![Figure 2: Distribution of questions based on their Socratic types](image-url)
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 $D \approx 110K$ (question, context) pairs from which about $L \approx 3.1K$ 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. 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

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

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). For QG experiments, we use the implementations provided by the transformers library for GPT-2 and T5 and ProphetNet. On our experimental machine, all models take between 12 – 19 hours for training depending on the task, specific model and the dataset used.

4.1 Classification Results

We studied several configurations based on BERT for training our question-type classifier. Table 2
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).

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

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; Svikhushina 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.

---

15Further details are included in Appendix C

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

![Figure 4: Distribution of “most likely to ask” questions chosen by crowdworkers](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Flu</th>
<th>Rel</th>
<th>Ans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>4.287</td>
<td>3.883</td>
<td>0.330</td>
</tr>
<tr>
<td>Worst</td>
<td>4.220</td>
<td>3.737</td>
<td>0.353</td>
</tr>
</tbody>
</table>

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). For this 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 the benefit of all. But the denial of voting rights to convicted felons after the end of their 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 SocraticQ. 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 antimicrobial 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 dial-

6 Conclusions
We created a novel dataset SocratiQ to support re-
s

search on automatic Socratic Question Generation.

We applied latest research in prompt-based condi-
tional 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 en-
abling future research on models for SoQG as well
as impactful applications in areas such as counsel-
ing 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
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 refer-
ence questions) in only 37-40% of the cases. Fur-
thermore, the question-type identification of auto-
mated 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 multturn data and also deduce via forum votes and
other indicators if the discussion indeed resulted
in changed minds and enabled alternate perspec-
tives. Furthermore, considering the special purpose
of Socratic questions in shaping perspectives and
enabling introspection and reflection, a compre-
prehensive 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 pre-
trained language models (PLMs) are readily avail-
able. 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 conclu-
sions or recommendations expressed in this mate-
rial 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 im-
proving 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


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


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


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


Supplementary Materials

A Composition of SocratiQ

We apply Latent Dirichlet Analysis using the topic modeling toolkit, Mallet (McCallum, 2002).\(^1\) 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, Racism, and Religion.

<table>
<thead>
<tr>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

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.

\(^1\)https://mallet.cs.umass.edu

Figure 5: Distribution of predicted question types in the SocratiQ

Figure 6: Confusion matrix of best classifier’s predictions on the test set
Table 9: Sample lexical cues and regular expression patterns used to tentatively map questions to their Socratic question types.

<table>
<thead>
<tr>
<th>Question type</th>
<th>Lexical Cues</th>
<th>Regex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarification</td>
<td>What do mean by ...?/What is the meaning of ...?</td>
<td>[Ww]hat[\s\w]+mean</td>
</tr>
<tr>
<td></td>
<td>How is ... related to ...?How are they related?</td>
<td>[Hh]ow[\s\w]+relate</td>
</tr>
<tr>
<td>Probing Assumptions</td>
<td>Are you assuming that...?Why are you assuming that...?</td>
<td>[Aa][e][s][m][u][n][t][i][o][n][g][+]</td>
</tr>
<tr>
<td></td>
<td>What is the assumption from...</td>
<td>[Ww]hat[\s\w]+assumption</td>
</tr>
<tr>
<td>Probing reasons and evidences</td>
<td>Where is your evidence that...?</td>
<td>[Ww]here[\s\w]+evidence</td>
</tr>
<tr>
<td></td>
<td>Why do you think...?</td>
<td>[Ww]hy[\s\w]+think</td>
</tr>
<tr>
<td>Probing implications and consequences</td>
<td>Are you implying...?</td>
<td>[Aa][e][m][i][p][l][i][y][ed]</td>
</tr>
<tr>
<td></td>
<td>What happens...?What would have happened if...?</td>
<td>[Ww]hat[\s\w]+happen</td>
</tr>
<tr>
<td>Probing alternative viewpoints and perspectives</td>
<td>What else...?</td>
<td>[Ww]hat[\s\w]+else</td>
</tr>
<tr>
<td></td>
<td>What other...?What are the other...?</td>
<td>[Ww]hat[\s\w]+other</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Qseq attr</th>
<th>Cseq attr</th>
<th>Cohen’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>.118</td>
<td>.033</td>
<td>1.59**</td>
</tr>
<tr>
<td>Clarification</td>
<td>.148</td>
<td>.041</td>
<td>1.85**</td>
</tr>
<tr>
<td>Probing reasons and evidences</td>
<td>.125</td>
<td>.037</td>
<td>1.64**</td>
</tr>
<tr>
<td>Probing implication and consequences</td>
<td>.108</td>
<td>.032</td>
<td>1.51**</td>
</tr>
<tr>
<td>Probing assumptions</td>
<td>.087</td>
<td>.020</td>
<td>1.76**</td>
</tr>
<tr>
<td>Probing alternate viewpoints and perspectives</td>
<td>.110</td>
<td>.025</td>
<td>1.80**</td>
</tr>
<tr>
<td>Others</td>
<td>.131</td>
<td>.061</td>
<td>0.76**</td>
</tr>
</tbody>
</table>
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” 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 learning rate scheduler to reduce the learning rate by a default factor of 0.1 whenever the F1-score from the validation set stops improving 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 the “gpt-large”, “t5-large”, and “Prophetnet-En” 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 and huggingface 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 except the learning rate, batch size, 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). For BLEURT (Sellam et al., 2020), we use their recommended BLEURT-20 model. Similarly, we use the recommended “microsoft/deberta-xlarge-mnli” 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.
<table>
<thead>
<tr>
<th>Clarification</th>
<th>Probing assumptions</th>
<th>Probing reasons and evidences</th>
<th>Probing implications and consequences</th>
<th>Probing alternative viewpoints and perspectives</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does that definition mean? (0.467)</td>
<td>How can one conclude... (0.439)</td>
<td>Why did that matter? (0.461)</td>
<td>Should we not care? (0.441)</td>
<td>Or what about drinking? (0.441)</td>
<td>Did you not listen? (0.431)</td>
</tr>
<tr>
<td>What do they mean? (0.457)</td>
<td>Is your hypothesis that... (0.385)</td>
<td>Was there a reason? (0.412)</td>
<td>Would it cause people... (0.434)</td>
<td>How about other physical... (0.434)</td>
<td>Who even cares about... (0.416)</td>
</tr>
<tr>
<td>Do you mean doctors? (0.441)</td>
<td>Would you have assumed... (0.376)</td>
<td>Can you list any... (0.408)</td>
<td>Would the situation change... (0.429)</td>
<td>What other species would... (0.424)</td>
<td>Do you not realise... (0.391)</td>
</tr>
<tr>
<td>How does that equal? (0.430)</td>
<td>What makes you presume... (0.360)</td>
<td>What level of evidence... (0.357)</td>
<td>Are you implying rich... (0.426)</td>
<td>What was there anything else... (0.415)</td>
<td>Did you not understand... (0.370)</td>
</tr>
<tr>
<td>How does this agree? (0.413)</td>
<td>Why does society assume... (0.338)</td>
<td>What term long evidence... (0.354)</td>
<td>What happens when inclusion... (0.413)</td>
<td>How would any other... (0.367)</td>
<td>What kind of dumbass... (0.347)</td>
</tr>
</tbody>
</table>

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

**CONTEXT:** A child should hold no religious (or absence 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:**
1. (Implications/Consequences) What do you think happens to young children who are forced to hold extremist views?
2. (Clarification) How does teaching kids about the bible relate to holding religious views?
3. (Assumptions) Why do you assume that adults would just move on from their parents when they are adults?
4. (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:**
1. (Implications/Consequences) Are you implying the bill could be identical between an adult and a child who are not of the same age?
2. (Clarification) What part of the point would you have that will encourage adults to move on?
3. (Reasons/Evidence) And why should a person be required to move out at all?
4. (Assumptions) Why do you assume that adults would just move on from their parents when they are adults?

**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.

**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:**
1. (Implications/Consequences) Are you asking the college cheerleaders to do something about it?
2. (Clarification) How do cheerleaders train on a surface that hard?
3. (Reasons/Evidence) Do you have any evidence to support the idea that cheerleaders in colleges might be underprivileged?
4. (Assumptions) Are we assuming cheerleaders are not working out to their maximum potential?

**Table 12:** Examples from SocratiQ are shown with their reference questions and questions generated with the ProphetNet−p model
For a question in a given context, label with the most-suitable question class

This qualification test will get you used to

- Clarification questions: Question probing to address any ambiguity from the context.
- Probing Assumptions: Question probing (for) any underlying assumption that supports the context.
- Probing Reasons and Evidence: Question probing (for) any justifications or concrete evidences that could have supported the context.
- Probing Alternatives viewpoints and Perspectives: Question probing (for) viewpoints alternate to those from the context.
- Probing Implication and Consequences: Question probing (for) any impacts or implication following a context.
- OTHER: Question not related to any of the Socratic question types above (e.g. rhetorical questions, irrelevant questions, illogical questions, etc.)

The following are some examples for each question class:

Clarification - Example 1

- [Context] If you actually want to change my view, you’ll have to show me more then a post with some shit talk. The two men in the “Biiiitchhh” Key and Peele skit pretend to call there wives bitches, but in reality are respectful to them.
- [Question] What kind of talk would you consider misogynistic?

Clarification - Example 2

- [Context] Disclaimer: I’m not trying to win this argument, I have my mind open on this, please change it. Who’s the scientist in my example? The person that says, “I don’t know how the sun rises?” They’re not wrong. They actually

Qualification Q2:

[Context]: Just to clarify, I’m not saying you have to accept everything (s)he does, like you can protest against their decisions if you want. In fact I support peaceful protests. You can voice your opinion as much as you want. I’m just saying that if you heap hatred on top of presidents in general, their job gets hard. Making the president’s job difficult makes decisions more difficult to make because (s)he knows that they will get more hatred towards one or the other party. And if the president fails, so does America and the rest of the world because they have our currency. Like for example, the current US president.

[Question]: If so, what happens when we have a President who doesn’t want what’s best for us?

Which is the most-suitable class to this question?

- Clarification
- Probing reasons and evidence
- Probing alternative viewpoints and Perspectives
- Probing Implications and consequences
- Probing Assumptions
- Others

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