How Well Do You Know Your Audience? Toward Socially-aware Question Generation

Ian Stewart and **Rada Mihalcea**
Computer Science and Engineering
University of Michigan
{ianbstew,mihalcea}@umich.edu

Abstract

When writing, a person may need to anticipate questions from their audience, but different social groups may ask very different types of questions. If someone is writing about a problem they want to resolve, what kind of follow-up question will a domain expert ask, and could the writer better address the expert's information needs by rewriting their original post? In this paper, we explore the task of *socially-aware* question generation. We collect a data set of questions and posts from social media, including background information about the question-askers' social groups. We find that different social groups, such as experts and novices, consistently ask different types of questions. We train several text-generation models that incorporate social information, and we find that a discrete social-representation model outperforms the text-only model when different social groups ask highly different questions from one another. Our work provides a framework for developing text generation models that can help writers anticipate the information expectations of highly different social groups.

1 Introduction

Writers are often expected to be aware of their audience (Park, 1986) and to minimize the effort required for others to understand them, especially if they cannot receive immediate feedback. However, NLP tools for writing assistance are not often made aware of the *social* composition of the audience (Ito et al., 2019; Zhang et al., 2020) and the information needs that different people may have. Preemptive writing feedback may therefore fail to help writers address the expectations of different people in their audience. This is especially important when the writer requests feedback from a specific group of people: in one post on a forum related to personal finance, a writer asks for help from financial "gurus" for advice about accepting a job offer.

A system that can preempt the hypothetical audience's information needs would enable the writer to revise their original post and avoid possible information gaps (Liu et al., 2012). Some online forums have already implemented crude solutions for this problem with automated reminders for writers to include basic information (e.g., location) in their post. Providing writers with preemptive questions can help especially in domains where different social groups have diverse information expectations. In the earlier example about personal finance, the advice-seeker could adapt their original post with answers to hypothetical "expert-level" questions (e.g. "Have you saved enough money for retirement?"), adding extra information that would enable experts to provide advice more quickly.

We cannot predict everyone's information needs, but some social groups with similar backgrounds (e.g., domain experts) will likely have consistent patterns in information expectations (Garimella et al., 2019; Welch et al., 2020). In this work, we evaluate several *socially-aware* question generation models with the goal of providing customized clarification questions to writers.

Our work contributes answers to the following questions:

- **How different are social groups based on the questions that they ask?** We collect a dataset of 200,000 Reddit posts seeking advice about a variety of everyday topics such as technology, legal issues, and finance, containing 700,000 questions.[1] We define several social groups that are relevant to possible information expectations such as expertise (§ 4.1). We demonstrate that different social groups, e.g. experts vs. novices, ask consistently different questions (§ 4.2, § 5.2.2).
- **How well can generation models predict socially-specific questions?** We extend an

[1] We will release the IDs for the post and author data, as well as the data processing code, to aid replication.

Proceedings of the SIGdial 2022 Conference, pages 255–269
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

existing generation model to incorporate social information about the question-askers (§ 5.1). In automated evaluation, a token-based socially-aware model outperforms the baseline for questions that are "divisive" and questions that are specific to a social group, particularly with respect to location as a social group (§ 5.2.3, § 5.2.4).

• **Are socially-aware questions useful for writers?** In human evaluations, we found that the socially-aware model is preferred over the text-only model for questions related to the question-asker's location and within the general advice-seeking domain (§ 5.3). This reinforces the utility of socially-aware models in scenarios where the social information is well-defined and where the topics are related to everyday concerns.

Importantly, the research presented in this paper shows that there are significant differences across groups with respect to questions they ask, and that we can develop models that are more attuned to these differences. Note that the goal of our work is not to improve the overall accuracy of a question generation system, but rather to develop methods that are sensitive to the needs of specific groups, thus paving the way toward technology that is available and useful for all.

2 Related Work

Question generation Question generation (QG) is unique among text generation tasks because it tries to address what a person *does not know*, rather than what they already know and want to write. QG systems are expected to create fluent and relevant questions based on prior text, in order to provide QA systems with augmented data (Dong et al., 2019) and students with question prompts to help their learning (Becker et al., 2012; Liu et al., 2012). In addition to typical supervised learning approaches (Du et al., 2017), reinforcement learning has proven useful, where questions are assigned a higher "reward" if they are more likely to have interesting answers (Qi et al., 2020a) and more relevant to the context (Rao and Daumé, 2019). Furthermore, work such as Gao et al. (2019) has proposed *controllable* generation techniques to encourage less generic questions, e.g. with higher difficulty. Such controllable-generation systems often leverage human-generated questions from a variety of domains, including Wikipedia (Du and Cardie, 2018), Stack Overflow (Kumar and Black, 2020), and Twitter (Xiong et al., 2019).

To our knowledge, prior work in question generation did not leverage the prior expectations of the question-askers. While sometimes providing controls for difficulty, no datasets currently include information about the inferred *background* of the question-askers. It seems natural that a person's prior knowledge would shape the information that they seek in response to a particular situation, yet analysis of the impact of social information on question generation remains absent. This study tests the role of social information in question generation using a dataset of posts from online forums, which feature complicated scenarios that can result in different information expectations between social groups.

Language model personalization Personalized language modeling often seeks to improve the performance of common language tasks, such as generation, using prior knowledge about the text's author (Paik et al., 2001). Personalization can improve task performance and make language processing more human-aware (Hovy, 2018), which ensures that a more diverse population is included in language models (Hovy and Spruit, 2016). To represent the text writer, personalized systems often integrate a writer's identity (Welch et al., 2020) or a writer's social network information (Del Tredici et al., 2019) into existing language models. A more generalizable approach converts the text-writer to a latent social representation such as an embedding (Pan and Ding, 2019), to be combined with the language representation in a neural network model where the social and text representations are learned jointly (Miura et al., 2017). We draw inspiration from the *contextualized* view of personalization from Flek (2020), and we represent the question-askers based on their prior behavior with respect to the specific *context* of a given post.

3 Data

In this study, we consider the task of generating clarification questions on information-sharing posts in online forums. We choose to study subreddits that have a high proportion of text-only posts, diverse topics, and where community members often ask information-seeking questions: `Advice` (lifestyle improvement), `AmItheAsshole` (social norms in complicated

Total posts	270694
Total questions	730620
Post length	304 ± 221
Question length	13.9 ± 8.08
Questions with question-asker data	77.7%
Questions with discrete question-asker data	75.2%
Questions with question-asker embeddings	43.5%

Table 1: Summary statistics about posts, questions, and question-asker data.

Subreddit	Posts	Questions
`Advice`	48858	87592
`AmItheAsshole`	61857	331345
`LegalAdvice`	53577	92737
`PCMasterRace`	31657	47613
`PersonalFinance`	74745	171333

Table 2: Summary statistics about subreddits.

situations), `LegalAdvice` (law disputes), `PCMasterRace` (computer technology), and `PersonalFinance` (money and investment). We collect all submissions ($\sim$ 8 million) to the above subreddits from January 2018 through December 2019, using a public archive (Baumgartner et al., 2020). We filter the post data to only include submissions written in English with at least 25 words, which we chose as a cutoff for posts that lack the context necessary for people to ask informed questions. To identify potential clarification questions, we collect all the comments of the submissions ($\sim$ 6 million) that are not written by bots, based on a list of known bot accounts like `AutoModerator`.

We conduct extensive filtering to include questions that are relevant and that seek extra information from the original post. The details are available in Appendix A. We summarize the overall data in Table 1, and we show the distribution of the posts and questions among subreddits in Table 2. Example posts and associated clarification questions are shown in Table 3.

4 Defining social groups

In this work, we assess the relevance of the question-asker's background in the task of question generation, by defining social groups and assessing their differences in question-asking.

4.1 Defining social groups

We collect a limited history for the question-askers ($N = 1000$ comments) to quantify relevant aspects of their background that may explain their information-seeking behavior. We consider the following social groups who are likely to have different information expectations:

1. `EXPERTISE`: A question-asker with less experience may ask about surface-level aspects of the post, while someone with more expertise might ask about a more fundamental aspect of the post. We quantify "expertise" using the proportion of prior comments that the question-asker made in the subreddit s (or a topically related subreddit; see § B.1) in which the original post was made. For example, if a question-asker has frequently written comments in `WallStreetBets` before asking a question in `PersonalFinance`, they are likely more familiar with financial terms than the average person. We define an `Expert` question-asker as anyone at or above the 75^{th} percentile of rate of commenting in a relevant subreddit, and a `Novice` question-asker as anyone below the percentile, where we chose the threshold to fit the skewed data distribution. Other threshold values produced similar results in social group classification.
2. `TIME`: A question-asker who replies soon after the original post was written may ask about missing information that is easily corrected (e.g. clarifying terminology), while a question-asker who replies more slowly may ask about more complicated aspects of the writer's request (e.g. the writer's intent). We quantify this with the mean speed of responses of the question-asker's prior comments *relative* to the parent post. We define a `Slow` question-asker as anyone at or above the 50^{th} percentile of mean response time, and a `Fast` question-asker as anyone below the threshold.
3. `LOCATION`: A question-asker who is based in the US may ask questions that reflect US-centric assumptions, while a non-US question-asker may ask about aspects of the post that are unfamiliar to them. We quantify location with the question-asker's self-identification from prior comments, using Stanza's English NER tool (Qi et al., 2020b) to identify LOCATION entities and OpenStreetMap to geo-locate the most likely locations. For those without self-identification, we identify all location-specific subreddits $\mathcal{S}_L$ in a question-asker's previous posts based on whether the subreddit name can be geolocated with high confidence (e.g., `r/NYC` maps to New York City). A question-asker a's location is identified with the location-specific subreddit where a writes at least 5 comments and where they write the most comments out of all location-specific subreddits $\mathcal{S}_L$.

Group category	Description	Social group	Example question	Example post title
EXPERTISE	Prior rate of commenting in the target subreddit, or a topically-related subreddit.	Expert ($\geq$ 75th percentile)	How much would you need to make on day 1 to meet your current financial obligations?	(PersonalFinance) Changing careers at 39
		Novice ($<$ 75th percentile)	Where do you live?	
TIME	Mean amount of time elapsed between original post and question-asker's comment, among all prior comments.	Fast ($<$ 50th percentile)	Does your wife have a relationship with him?	(LegalAdvice) Having a child and partner's father is sex offender
		Slow ($\geq$ 50th percentile)	If he is a sex offender, shouldn't he be kept away from children?	
LOCATION	Inferred location of question-asker.	US	Have you looked at the RX 580?	(PCMasterRace) Should I buy GTX 1050Ti?
		non-US	The 1050 is 160$ in India?	

Table 3: Group categories for question-askers, with example questions and posts.

Group category	Top-3 LIWC categories (absolute frequency difference)
LOCATION	
US >non-US	MONEY (0.512%), WORK (0.361%), RELATIV (0.337%)
non-US >US	FOCUSPRESENT (0.356%), FUNCTION (0.327%), AUXVERB (0.305%)
EXPERTISE	
Expert >Novice	MONEY (0.207%), YOU (0.135%), FOCUSPRESENT (0.106%)
Novice >Expert	DRIVES (0.097%), AFFILIATION (0.056%), REWARD (0.055%)
TIME	
Fast >Slow	YOU (0.312%), PPRON (0.225%), PRONOUN (0.160%)
Slow >Fast	DRIVES (0.105%), AFFECT (0.082%), IPRON (0.066%)

Table 4: LIWC category word usage differences across social groups (% indicates absolute difference in normalized frequency). All differences are significant with $p < 0.05$ via Mann-Whitney U test.

We summarize these definitions of different social groups in Table 3. The example questions in demonstrate that question-askers who occupy different groups tend to ask questions about different aspects of the original post: e.g. the Fast question-asker addresses a basic fact about the situation, while the Slow question-asker addresses a more complicated/hypothetical point.

4.2 Validating group differences

As a first step, we test for consistent differences in the types of questions asked by different social groups. We test for topical differences between the groups by comparing the relative rate of LIWC word usage in their questions, a common strategy to identify salient differences between social groups (Pennebaker et al., 2001). The results in Table 4 show consistent differences in word usage in the questions. Expert question-askers ask about money more often than Novices, which could indicate an assumption from prior experience that post authors' core problems stem from their financial decisions (even outside of the finance-related subreddits). Similarly, US question-askers have more questions about money and work than non-US readers, who often frame questions to address present-tense issues and write with more auxiliary verbs. Fast-response question-askers ask more often about the post author (YOU), which may indicate a stronger interest toward the post author's background, as opposed to slow-response question-askers who address the poster's high-level intentions (DRIVE) and emotional behavior (AFFECT). While it is possible that some of these differences are spurious, it is unlikely that they all relate to stylistic patterns such as regional differences (LOCATION), considering the prevalence of relevant LIWC categories (e.g. MONEY relates to financial questions, which are relevant to the data).

We verify these differences with a classification task, which we detail in § B.2.

5 Question generation

5.1 Model design

We build the generation models on top of the BART model (Lewis et al., 2020), a transformer model known to be resistant to data noise. We use the same pre-trained model (bart-base; $|V| \approx 50,000$) and the same training settings for all models.[2] The main point of the model modifications is not to achieve universally high accuracy, but to assess the value of different social data representations in question generation.

[2]Learning rate 0.0001, weight decay 0.01, Adam optimizer, 10 training epochs, batch size 2, max source length 1024 tokens, max target length 64 tokens, cross-entropy loss.

5.1.1 Social tokens

For the "social-token" model (a discrete representation), we add a special token $\{GROUP_g\}$ to the text input of the baseline model to indicate whether the asker belongs to social group g (cf. prior work in controllable generation; Keskar et al. 2019). The embeddings for these social tokens are learned during training in the same way as the other text tokens. All question-askers who could not be assigned to a group are represented with `UNK` tokens.

5.1.2 Social attention

For discrete modeling, we also consider customizing a separate part of the model for different social groups. Specifically, we change one of the attention layers of the typical transformer model (Vaswani et al., 2017) to represent differences in how different question-askers may perceive a post.

We replace attention module ℓ in the encoder with a different module for each social group g. For regularization, we train a separate *generic* attention module at the same time as the social-group attention, concatenate the social attention with the generic attention, and pass the concatenated attention through a linear layer to produce the final attention distribution. We choose the layer index $\ell = 1$ from $\{1, 3, 5\}$ through performance on validation data. For a question written by an asker who belongs to group g (gen indicates generic attention, f indicates a feed-forward linear layer), the attention is computed as follows:

$$\text{Multihead}_\ell(x) = f([\text{Multihead}_g(x); \text{Multihead}_{gen}(x)])$$

5.1.3 Social embeddings

For a *continuous* approach to personalization (Wu et al., 2021), we represent question-askers using latent embeddings $e^{(a)}$ based on their prior *subreddit* and *text* behavior.

For *subreddit* behavior, we compute the cross-posting matrix $\mathcal{P}$ for all subreddits and all question-askers in our data, where $P_{i,j}$ is equal to the NPMI of question-asker j writing a comment in subreddit i. We compress the matrix using SVD ($d = 100$), and the subreddit embedding $e_s^{(a)}$ for question-asker a is set to the average of the embeddings across all subreddits in which a previously posted. For *text*, we compute an embedding based on the question-asker's previous comments. We train a `Doc2Vec` model $\mathcal{D}$ (Le and Mikolov, 2014) on all prior comments and represent each comment as a single document embedding ($d = 100$, default skip-gram parameters). The text embedding $e_t^{(a)}$ for question-asker a is computed as the average over all prior comments.

To add the social embedding to the input text, we pass $e^{(a)}$ through a linear layer to match the text dimensionality ($d = 768$). We append a special "social embedding" token and the embedding $\hat{e}^{(a)}$ to the end of the text input.

5.2 Results

We use the models proposed above and a text-only baseline, and train them on the same task of question generation. We use a sample of our data for training/testing, for a total of 155396 questions for training, 51774 for validation, and 53080 for test.

We use the following metrics to automatically evaluate text quality for target question q and generated question $\hat{q}$: BLEU-1 (single word overlap between q and $\hat{q}$); perplexity; BERT Distance (cosine distance between sentence embeddings for q and $\hat{q}$, via the same DistilBERT system used throughout; Sanh et al. 2019); Type/token ratio (among bigrams in $\hat{q}$); Diversity (% unique questions among all generated questions $\hat{\mathcal{Q}}$); Redundancy (% generated questions $\hat{\mathcal{Q}}$ that also appear in training data $\mathcal{Q}_{\text{train}}$). The text overlap metrics like BLEU are important in judging performance even in our open-domain setting, because the models should produce questions that are faithful to the original intent of the question-askers (Wu et al., 2021). Without measuring overlap, it would be possible for a socially-aware model to generate highly diverse questions that are completely unrelated to the question-asker's intent.

5.2.1 Aggregate results

The aggregate results are shown in Table 5. Overall, we see that the simpler socially-aware models (tokens and attention) perform roughly the same as the text-only model via traditional BLEU and BERT Distance metrics. The socially-aware model generates questions that have higher overall diversity, but also higher perplexity. These results echo prior work in text generation which finds that models which incorporate pragmatic information often produce more diverse text than expected (Schüz et al., 2021). The higher perplexity can be explained

stat	BLEU-1 ↑	BERT Dist. ↓	Diversity ↑	Type/token ↑	Redundancy ↓	PPL ↓
Text-only	**0.159**	**0.728**	0.613	0.122	**0.187**	**264**
Social token	0.159	0.731	0.675	**0.127**	0.191	271
Social attention	0.157	0.752	0.511	0.068	0.468	488
Subreddit embedding	0.153	0.746	**0.744**	0.091	0.277	657
Text embedding	0.154	0.745	0.732	0.090	0.292	609

Table 5: Question generation results by model on full test data. ↑ means higher score is better, ↓ means lower is better.

Subreddit	LegalAdvice	AmITheAsshole
Text context	My five year old son is in kindergarten. The teacher let the kids out of their recess area and did not watch them properly, and my son got lost.	My roommate has been dating someone with a young child. Both the woman and her child are generally annoying.
Social group	LOCATION (US)	EXPERTISE (Novice)
Actual question	What is your location?	Have you talked to your roommate?
Text-only	What are your damages?	Have you spoken to your roommate about this?
Social token	Was this a private school or a government agency?	Do you and your roommate pay rent to the landlord?
Model performance	social token > text-only (BERT Dist.)	text-only > social-token (BLEU)

Table 6: Example posts, target questions, and generated questions.

partly by the unconstrained nature of the generation task (e.g., not providing an answer to generate the question; see § 6.2) as well as the relatively complex nature of most of the questions.

5.2.2 Qualitative analysis of model output

We first show several examples of generated text (Table 6). In a legal context (first column), the social-token model correctly predicts that the question-asker will focus on the location of the incident rather than the outcome (text-only model), possibly because a US question-asker may have location-specific advice to provide.

We also use the social-token model to generate attention distributions over the input sequence for different groups. We input the same text for both reader groups in the same category, changing only the social token appended to the text. We compute the attention distribution from the first layer of the encoder, compute per-word attention scores via the mean over all heads and all token-pairs, and compute the ratio of attention for each group category. The distributions for an example post are shown in Table 7, and they seem to match our earlier findings with word category differences (§ 4.2). For LOCATION, we see that the model prompted with a US token pays more attention to MONEY words ("booked tickets"), while the model prompted with NONUS focuses on time-related words that could be translated to FOCUSPRESENT in the question ("happened," "'few days ago"). For Expertise, the NOVICE social token produces higher attention on social relationships ("friend," "daughter"), and the model with EXPERT input attends to pronouns that could be converted to "you" pronouns in a following question ("my"). For TIME, the model with SLOW input pays attention to DRIVE words ("planning," "looking"), while the model with FAST input pays more attention to personal pronouns ("I", "my"). While we do not perform large-scale annotation of attention distributions, the examples shown here complement the generated text and reveal potential concepts that the model has learned to associate with different social groups.

5.2.3 Divisive posts

Socially-aware question generation should perform well in cases where different social groups have divergent opinions, e.g. where experts disagree with novices. We now test the models' ability to predict divisive questions. For post p, question q_1 written by an author of group 1, and question q_2 written by an author of group 2, we define $\text{sim}(q_1, q_2)$ based on the cosine similarity of the latent representations of the questions, generated by DistilBERT as before (Sanh et al., 2019). We label as "divisive" all pairs of questions that have a similarity score in the lowest n^{th} percentiles. We show examples of divisive posts in § C.3.

The results of the question prediction task on divisive posts are shown in Table 8. The social-token model slightly outperforms the text-only baseline for questions that are highly dissimilar (i.e. less similar than 90% and 95% of the question pairs), and all socially-aware models tend to do better in diversity. This suggests that the social-token model may pick up information specific to the different social groups that is required to anticipate how the question-askers approach potentially subjective posts. We also note the unusually high perplexity across all models,

EXPERTISE EXPERT, NOVICE	So my friend is having difficulty getting her 15 year old daughter to school . My friend will let her off at school , watch her enter the building , and then later will find her back home during school time .
LOCATION NONUS, US	This happened a few days ago and my friend thought I was a bit rude , but I felt I was totally justified . So we booked tickets for a nearly full flight and the only row with 2 seats beside each other had somebody that already booked the seat...
TIME SLOW, FAST	Folks , I am planning to return to PCs after an absence . my budget is about 3 k and I already found a machine that will be around 2 , 5 k . So right now I am searching for monitors and I am looking for...

Table 7: Ratio of encoder attention generated by social-token model for input conditioned on different social groups. Attention computed via mean over all pairwise scores between tokens.

Model	BLEU-1	Div.	Red.	PPL
sim(q_1, q_2) $\leq$ 5% (N=1074)				
Text-only	0.137	0.688	0.222	383
Social token	**0.142**	0.771	**0.208**	**359**
Social attention	0.130	**0.875**	0.479	601
Subreddit emb.	0.137	0.854	0.292	945
Text emb.	0.137	0.840	0.375	623
sim(q_1, q_2) $\leq$ 10% (N=2146)				
Text-only	0.160	0.699	**0.232**	**325**
Social token	**0.164**	0.781	0.235	327
Social attention	0.155	0.798	0.500	547
Subreddit emb.	0.148	0.864	0.308	1048
Text emb.	0.150	**0.868**	0.348	617

Table 8: Question generation results for divisive posts.

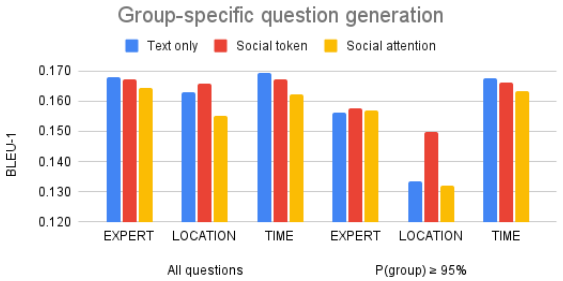

Figure 1: BLEU-1 scores for question generation, on (1) full data and (2) subset of data with high group-specific probability (determined by classifier).

which may indicate that socially-specific questions are complicated and far from "normal" questions.

5.2.4 Group-specific questions

We investigate another desired property of socially-aware models, the ability to predict questions that are strongly associated with a particular group. Post writers would benefit from such questions, e.g. technical questions from "expert" askers, because these questions would help the post writer preempt specific and unexpected information needs from that group. We subset the data to all questions q with question-asker a that the trained social group classifiers assign to the group g_a with high confidence ($P \geq 95\%$) (see § B.2 for classifier details).

We report the results for this data subset in Figure 1. The relative performance of the socially-aware models increases when only considering data with highly group-specific questions. This is particularly apparent for the LOCATION group category, illustrated by the following example. In our data, a socially-specific question was written by a non-US question-asker in LegalAdvice in response to a post about a mailing problem: "Have you sent a change of address notice to the post office?" In this situation, the social-token model generated the question "Did you give them your current address?" The social-token model seems to have identified a concern that a non-US question-asker might be more likely to focus on (e.g. due to moving frequently) than a US question-asker.

5.3 Human evaluation

To corroborate the generation results about divisive questions, we collect human annotations about: (1) question quality; and (2) guessing the social group based on the generated question (see § C.6 for (2)).

We use the text-only model and the social-token model to generate questions from a sample of the test data, as follows. For each subreddit s and social group category $\mathcal{G}$, we sample up to $N = 10$ posts that have divisive questions from groups g_1 and g_2 where the similarity is below the 10^{th} percentile (§ 5.2.3).[3] We then generate a single question for the post from the text-only model and two questions from the social-token model, one for each social group in the category (e.g., for EXPERTISE, q_1 for Expert and q_2 for Novice). We provide details

[3]Some combinations of subreddits and reader groups have fewer than 10 posts, due to the data sampling strategy.

Text type	A	R	U
Overall			
Ground-truth	3.83	3.59	3.92
Text-only	3.84	3.68*	3.96*
Social-token	3.80	3.35	3.73
Social group			
EXPERTISE			
Ground-truth	3.89	3.68	3.85
Text-only	3.81	3.62*	3.91*
Social-token	3.61	2.99	3.49
LOCATION			
Ground-truth	4.06	3.58	4.20
Text-only	4.01	3.69	4.05
Social-token	4.20	3.63	4.19
TIME			
Ground-truth	3.64	3.52	3.83
Text-only	3.77	3.74	3.95*
Social-token	3.74	3.53	3.70
Subreddit			
Advice			
Ground-truth	3.75	3.63	3.91
Text-only	3.32	3.29	3.57
Social-token	3.49	3.15	3.67
AmItheAsshole			
Ground-truth	3.79	3.58	4.01
Text-only	3.74	3.61	3.89
Social-token	3.82	3.39	3.69
LegalAdvice			
Ground-truth	4.18	3.88	4.47
Text-only	3.95	3.60	4.19*
Social-token	3.86	3.23	3.81
PCMasterRace			
Ground-truth	3.72	3.44	3.62
Text-only	4.20	4.07*	4.16
Social-token	3.98	3.39	3.84
PersonalFinance			
Ground-truth	3.72	3.43	3.58
Text-only	4.04	3.89	4.01*
Social-token	3.87	3.56	3.70

Table 9: Human annotation scores for question quality, including **A**nswerable, **R**elevant, **U**nderstandable (scale 1-5). * indicates that the score is greater than the scores from the other model type with $p > 0.05$ (Wilcoxon test). Underline indicates best generation model.

of annotation in § C.5.

We show the results in Table 9. The annotators in aggregate preferred the questions from the text-only model over the social-token model. However, the social-token model questions were perceived as more answerable and understandable for questions generated using LOCATION information, which aligns with prior results (§ 5.2.4). The social-token model is also perceived as more answerable and understandable in the context of Advice, which makes sense considering that the social-token model has more diverse output that may suit the broad domain of general-advice posts.

We show example generated and actual questions with their human evaluation ratings in Table 10. In the first example, the text-only model addresses an important missing gap in original post (how the scammer got information), while the social-token model seems to focus too much on details ("on the hook") which leads to a less understandable question. In the second example, the social-token model addresses missing information that may be more salient to a non-US question-asker who wants to know more about homelessness (possibly less salient to a US question-asker), while the text-only model produces a question that is not answerable due to a misunderstanding of the original post (focus on the text rather than the writer). Note that this type of question is not marked by a surface-level feature such as regional style, but rather a deeper focus on cause and effect, which suggests that the model has learned more fundamental differences about the nature of LOCATION as a social group.

Subreddit	LegalAdvice	Advice
Text context	My mother lost $50000 on an online dating site to a scam. If something happened to her, would I be on the hook for this?	I want to break up with my girlfriend but: number 1 I don't want to hurt her, number 2 I don't know if I can manage on my own, number 3 I don't always believe in myself, and if I lose my job I'll be homeless.
Social group	EXPERTISE (Expert)	LOCATION (non-US)
Actual question	Has your mother contacted the police? (Understandable=4.67)	Have you tried talking to her? (Answerable=5)
Text-only model	How did the scammer get the info from your Mom? (Understandable=4.33)	Number 2 doesn't even sound like a good idea, have you tried number 3? (Answerable=2.33)
Social token model	Are you on the hook for what? (Understandable=2.67)	Why do you think you'll be homeless? (Answerable=5)

Table 10: Example questions with human evaluation scores.

6 Conclusion

6.1 Discussion

This study evaluated the incorporation of social information into question generation, to help writers understand the information needs of different people. We found that social groups related to expertise, time, and location can all be differentiated based on the questions that they ask. In generation, the discrete social representations outperformed continuous representations, and the social-token model outperformed the baseline when the questions are divisive. In human evaluation, the social-token models produced better output for the LOCATION group, implying a more clear definition versus other social groups.

Future research in question generation should focus on divisive questions as the main area of improvement. Researchers may also consider ensemble models (Liu et al., 2021) that use a text-only generator with less subjective input text (e.g., in technical settings), and a social-token generator in more divisive settings. For future evaluation, socially-aware question generation may benefit other contexts such as journalism, medicine, and public policy, where people are likely to have differing information needs based on their background experience (Assmann and Diakopoulos, 2017). No matter the case, writers will always benefit from knowing in advance what information their audience will need.

6.2 Limitations

The primary limitations of this work relate to the definition of "social group," which may have contributed to the minimal gains by the social token model. This work focused on generic social groups that can be extended to other domains, which may leave out domain-specific social groups (e.g., socioeconomic status). The social groups may not mean the same thing in different domains: an `EXPERT` question-asker in the legal domain may be a professional lawyer, while in personal advice the average `EXPERT` may lack professional experience. Most notably, the social groups used in this work were not validated by any annotators or by the question-askers. This especially matters for the `EXPERTISE` category, considering the subjective status of expertise within online communities (Johnson, 2001). To accurately identify non-obvious social groups, researchers should ask domain experts to label a small set of user data as gold labels, and then compare the automated labels against this gold standard set.

In terms of the task, this work focuses on unconstrained question generation, i.e. we do not use answers (Dong et al., 2019) or intentions (Cao and Wang, 2021) to guide generation. The results presented in this work represent a lower bound on performance, which includes unusually high perplexity (Table 5) and sometimes unexpected topic choices (Table 6). This problem is compounded by the fact that social group information may not always be useful e.g. for non-divisive questions, and therefore such social guidance may simply confuse the model. Future work would collect both questions and answers, or at least question type labels, to provide consistent guidance for socially-aware question generation.

7 Ethics statement

We acknowledge that text generation is an ethically fraught application of NLP that can be used to manipulate public opinion (Zellers et al., 2020) and reinforce negative stereotypes (Bender et al., 2021). Our models could be modified to generate abusive or factually misleading questions, which we do not endorse. Furthermore, our models may accidentally memorize private information from the training data. We intend for our work to benefit people who share information about themselves for the purpose of gaining feedback from peers.

All data used in this project was publicly available via the Pushshift API (Baumgartner et al., 2020). In our final release we will not release any data with personally identifiable information (e.g., `LOCATION` data), in order to protect the original authors. This is not ideal considering that `LOCATION` seemed to be the most useful input to the model, but the remaining social attributes may prove useful for future researchers who want to test other definitions of "divisive" questions, e.g. positive versus negative valence. Furthermore, we do not claim that we have the perfect definitions of the social groups that we attempted to identify in our study, and it is possible that a Reddit user who finds themselves labeled as e.g. an "expert" would disagree. We encourage future researchers to compare their own definition of the various social groups against our own labels, e.g. a different definition of "expertise."

Acknowledgments

This material is based in part on work supported by the John Templeton Foundation (grant #61156) and by the Michigan Institute for Data Science (MIDAS). Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or MIDAS. We thank members of the LIT Lab, including Artem Abzaliev and Aylin Gunal, for their help piloting the human annotation task, and for providing feedback on early results. We also thank the annotators who identified valid questions as part of the data filtering process.

References

Karin Assmann and Nicholas Diakopoulos. 2017. Negotiating change: Audience engagement editors as newsroom intermediaries. In *International symposium on online journalism (ISOJ)*, pages 25–44.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *ICWSM*, volume 14, pages 830–839.

Lee Becker, Sumit Basu, and Lucy Vanderwende. 2012. Mind the gap: Learning to choose gaps for question generation. In *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 742–751.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. 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 6424–6439.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4701–4711.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *NeurIPS*.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pages 1907–1917.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1342–1352.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838.

Yifan Gao, Lidong Bing, Wang Chen, Michael R. Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *IJCAI*, pages 4968–4974.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the Rough: Generating Fluent Sentences from Early-Stage Drafts for Academic Writing Assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 40–53.

Christopher M Johnson. 2001. A survey of current research on online communities of practice. *The internet and higher education*, 4(1):45–60.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *EACL*, pages 427–431.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Vaibhav Kumar and Alan W. Black. 2020. ClarQ: A large-scale and diverse dataset for Clarification Question Generation. In *ACL*, pages 7296–7301.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In *ACL*, pages 6691–6706.

Ming Liu, Rafael A. Calvo, and Vasile Rus. 2012. G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support. *Dialogue & Discourse*, 3(2):101–124.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1260–1272.

Woojin Paik, Sibel Yilmazel, Eric Brown, Maryjane Poulin, Stephane Dubon, and Christophe Amice. 2001. Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In *Proceedings of the 1st international conference on Knowledge capture*, pages 116–122.

Shimei Pan and Tao Ding. 2019. Social media-based user embedding: A literature review. In *IJCAI*.

Douglas B Park. 1986. Analyzing audiences. *College Composition and Communication*, 37(4):478–488.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020a. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 25–40.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020b. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Sudha Rao and Hal Daumé. 2019. Answer-based adversarial training for generating clarification questions. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:143–155.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *EC2*.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Charles Welch, Jonathan K Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. In *NeurIPS*.

Justine Zhang, James Pennebaker, Susan Dumais, and Eric Horvitz. 2020. Configuring audiences: A case study of email communication. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26.

A Data: question filtering

Initial analysis revealed that some questions were either irrelevant to the post (e.g., "what about X" where X is unrelated to the post topic) or did not actually seek more information from the original post (e.g., rhetorical questions). To address this, we sampled 100 questions from each subreddit in the data along with the parent post, and we collected binary annotations for relevance ("question is relevant") and information-seeking ("question asks for more information") from three annotators, who are undergraduate students and native English speakers. We provided instructions and a sample of 20 questions labeled by one of the authors as training data for the annotators. On the full data, the annotators achieved fair agreement on question relevance ($\kappa = 0.56$) and on whether questions are information-seeking ($\kappa = 0.62$).

After annotation, we removed all instances of disagreement among annotators to yield questions with perfect agreement for relevance (76% perfect-agreement) and information-seeking (80%). In the perfect-agreement data, the majority of questions (94%) were marked as relevant by both annotators, which makes sense considering that the advice forums generally attract good-faith responses from commenters. We therefore chose to not filter questions based on potential relevance. To filter information-seeking questions, we trained a simple bag-of-words classifier on the annotated data (binary 1/0; based on questions with perfect annotator agreement).[4] The annotated data were split into 10 folds for training and testing, and the model achieved 87.5% mean F1 score, which is reasonable for "noisy" user-generated text. We applied the classifier to the full dataset and removed questions for which the classifier's probability was below 50%.

B Defining social groups

B.1 Social embeddings: topically-related subreddits

In our discrete-representation models, the criterion for defining a question-asker for post p in subreddit s as an `Expert` or `Novice` is whether they have previously written comments in s or in a topically similar subreddit.

We find similar subreddits for each target subreddit s by (1) computing the top-20 nearest neighbors for subreddit s in subreddit embedding space (see § 5.1.3) and (2) manually filtering unrelated subreddits. We report the related subreddits in Table 11.

Subreddit	Neighbors
`Advice`	`answers, ask, askdocs, dating_advice, getdisciplined, mentalhealth, needadvice, socialskills, tipofmytongue`
`AmItheAsshole`	`askdocs, isitbullshit, tooafraidtoask`
`LegalAdvice`	`askhr, bestoflegaladvice, insurance, landlord, lawschool, legaladviceuk, scams`
`PCMasterRace`	`bapcsalescanada, buildmeapc, linuxmasterrace, monitors, overclocking, pcgaming, suggestalaptop, watercooling`
`PersonalFinance`	`accounting, askcarsales, churning, creditcards, financialindependence, financialplanning, investing, realestate, smallbusiness, studentloans, tax, whatcarshouldibuy, yna`

Table 11: Filtered neighbor subreddits for advice-related subreddits.

B.2 Validating group differences: classification

To verify the differences in question content observed in § 4.2, we train a single-layer neural network to classify social groups, using a latent semantic representation of the question-asker's question q and the related post p generated by the DistilBERT transformer model (Sanh et al., 2019). The embedding for the question and the post are each converted to $d = 100$ dimensions via PCA for regularization, and then concatenated. We train a separate model for each subreddit, and we

[4] We restricted the vocabulary to the 50 most frequent words, minus stop-words, to avoid overfitting. Initial tests with SVM, logistic regression, and random forest models revealed that the random forest model performed the best, which we used for the final classification model.

Features	Social group	Accuracy
Question text	EXPERTISE	70.1 (± 2.5)
	TIME	81.6 (± 7.5)
	LOCATION	75.4 (± 2.8)
Post + question text	EXPERTISE	73.5 (± 6.4)
	TIME	83.1 (± 8.3)
	LOCATION	66.3 (± 10.2)

Table 12: Social group prediction accuracy (mean, standard deviation measured across subreddits).

up-sample data from the minority class.

We report mean accuracy over all subreddits in Table 12. The models consistently outperform the random baseline across all group categories tested, which suggests a clear difference between social group members. The models trained on the combined post and question text generally help prediction improve over the question text alone, which supports the hypothesis that a question-asker's background is reflected in both the question they ask and the context in which the question is asked. Therefore, generating group-specific questions requires understanding how the question relates to the original post content, in addition to the question writing style. We find an unusually high performance for TIME, which may be due to a more consistent writing style among Fast question-askers.

C Results: question generation

We report here the results of additional tests to evaluate the relative utility of the socially-aware models with respect to different types of question-post scenarios.

C.1 Performance by question type

First, we assess the relative performance of different question generation models according to the type of question asked. Questions are categorized based on the root question word, e.g. "who," "what," "when."[5] We compare the BLEU-1 scores of all question generation models on the specified questions, restricting to questions asked by question-askers who could be assigned to at least one social group or an embedding.

The results are shown in Figure 2. In contrast to the aggregate results, the social-attention model outperforms the text-only baseline for "do," "where," and "who" questions. All socially-aware models outperform the text-only model for "when" questions. These questions may reflect more of a focus on concrete details such as locations, times, and people mentioned by the original poster, and therefore the socially-aware models may generally identify differences among question-askers in terms of the details requested. The text-only model outperforms the socially-aware models for questions that are potentially more subjective, including "can," "could," "would," and "should" questions. These more subjective questions may require the models to focus more precisely on the original post (e.g. a "would" question to pose a hypothetical concern about the post author's situation), and therefore such questions may be less dependent on question-asker identity.

C.2 Post similarity

A helpful question should be related to the original post, but should not be so similar that it requests information that the post has already provided. We therefore assess the tendency of the models to generate semantically related questions for the given posts. We compute the similarity between each generated question q and the associated post p using the maximum cosine similarity between the sentence embedding for q and each sentence s in p. The sentence embeddings are generated using the DistilBERT model (Sanh et al., 2019).

The results in Figure 3 show that the best overall models, text-only and social-token, generate questions that are more similar to the original post than expected (cf. "target text" i.e. ground-truth). The other socially-aware models show a significantly lower similarity, implying that their generated questions address *new information* about the post that is not mentioned in the post itself. For example, in response to a r/Advice post about self-improvement ("I just need some tips on maybe motivating myself"), the model with social text embeddings asks "What do you want to do with your life?" The generated question is less semantically similar to the original post than the target question ("Have you talked to a doctor about this?") but addresses an underlying personal issue for the post author that only a particularly thoughtful question-asker would uncover.

C.3 Divisive questions: examples

We provide examples of divisive posts in Table 13 (§ 5.2.3). For TIME, the Slow question-asker seems to target a more complicated and underlying

[5] We use the dependency parser from spacy (Honnibal and Johnson, 2015) to identify root question words based on their dependency to the root verb of the question (e.g. advmod for "where" in "where do you live?").

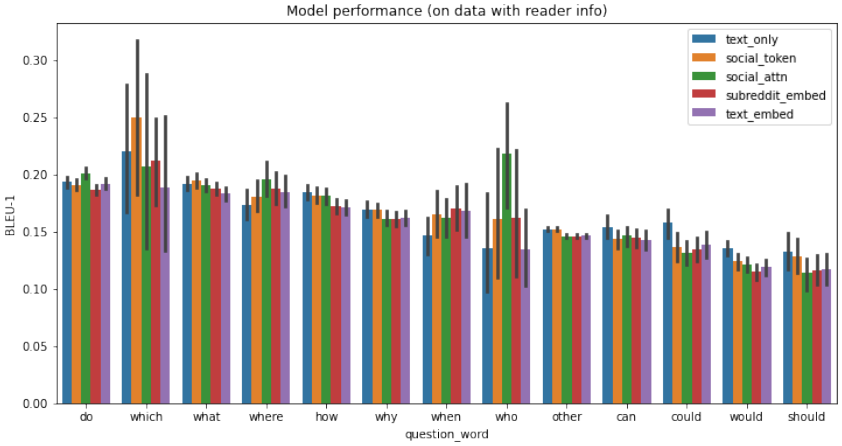

Figure 2: Model performance by question type.

Subreddit	PersonalFinance	LegalAdvice	AmITheAsshole
Text context	I need help figuring out what's the best next step. I have $1200 saved for car payments but I have no idea after that.	Last month I got a letter from a law firm representing someone that I owe a debt to. Two years ago I couldn't continue to make payments to the creditor and almost went bankrupt.	My younger brother is autistic. He can function and he has a job (janitor), hangs out with his friends but he can't live on his own.
Social group	EXPERTISE	TIME	LOCATION
Group 1	(Expert) Have you been applying for jobs all day?	(Slow) Have you asked what they are willing to settle for?	(US) What if down the road you had to re-locate for work or your wife's work?
Group 2	(Novice) Are you above water on the car?	(Fast) Do you actually intend on filing bankruptcy?	(non-US) How disabled is your brother?
Question similarity	0.209	0.256	0.190

Table 13: Example divisive questions for different social groups.

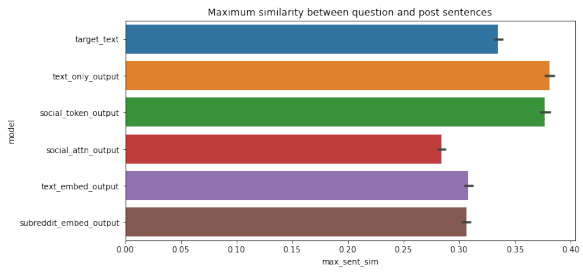

Figure 3: Maximum semantic similarity between questions and sentence from original post.

issue around the debt problem, while the Fast question-seeker clarifies a basic detail about the case. For LOCATION, the question from the US asker focuses on adapting to work needs, while the non-US question addresses the writer's brother and his medical situation. In all cases, we can see that these kinds of questions are more likely to be anticipated by a generation model that produces more diverse output.

C.4 Divisive posts: word embeddings

In § 5.2.3, we identified questions as "divisive" based on low similarity between the latent representations of the questions, as generated by a sentence encoder. We also experiment with determining divisiveness based on static word embeddings. We leverage a set of word embeddings trained with the FastText algorithm (Joulin et al., 2017), and we convert each questions to a latent representation using the average over all embeddings for the tokens in the question. We then compute paired question similarity as before, with cosine similarity. The questions from the sentence embeddings and those from the static word embeddings have a high degree of overlap: setting the similarity threshold below 5% yields an overlap of 23.7%, and a similarity threshold below 10% yields an overlap of 45.3%. Next, we test the correlation between the sentence embedding similarity and the word embedding similarity and find a high amount of correlation ($R = 0.98$, $p < 0.001$). We

Data	Accuracy
Overall	47.5
Social group	
EXPERTISE	49.3
LOCATION	60.8
TIME	36.9
Subreddit	
Advice	45.6
AmItheAsshole	48.9
LegalAdvice	53.3
PCMasterRace	42.9
PersonalFinance	47.8

Table 14: Human annotation accuracy for group guessing task.

conclude that labeling divisive questions using word embedding similarity rather than sentence embedding similarity would yield similar results to those observed earlier.

C.5 Human evaluation: annotation details

We provide the details of the annotation required for the human evaluation task (§ 5.3). We annotate the questions for each combination of subreddit and group category, and we recruit 1 annotator per task via Prolific, with 3 social groups $\times$ 5 subreddits $\times$ 3 annotators = 45 annotators total, and a maximum of 50 questions total for each annotator. For domain-specific subreddits, we recruit annotators based on profession, e.g. annotators who work in the finance industry for r/PersonalFinance. We pay our annotators $5 for the task, assuming about 30 minutes per task. Annotators judged question quality on a 5-point scale based on whether they were answerable, relevant, and understandable. The annotators achieved reasonable agreement considering the subjective nature of the task, with Krippendorff's alpha at 0.153 for "Answerable," 0.309 for "Relevant," and 0.23 for "Understandable" (compared to 0 for random chance).

C.6 Human evaluation: social group prediction

We report here the results of the additional annotation task mentioned in § 5.3. Following the question quality task, for each post we provide the two social-token model questions in random order for a group prediction task, where annotators must choose the question that corresponds to a given social group in the category: e.g. "Which question was more likely to be written by an **expert** reader?" We show the results for the group-guessing task in Table 14. Annotators generally had trouble guessing the identity of the social groups except for the LOCATION category, which corresponds with the higher quality ratings reported in Table 9. We also find slightly higher guessing accuracy for LegalAdvice, which may be due to intuitive understanding among annotators on what constitutes a difference in social groups for the legal domain (e.g. experts using particular terminology). The low performance in this task may indicate that human-understandable differences between the questions may be less obvious in individual pairs of questions as compared to the aggregate groups of questions (see differences in § 4.2).