What Types of Questions Require Conversation to Answer? A Case Study of AskReddit Questions

Shih-Hong Huang
The Pennsylvania State University
University Park, PA, USA
szh277@psu.edu

Chieh-Yang Huang
The Pennsylvania State University
University Park, PA, USA
chiehyang@psu.edu

Ya-Fang Lin
The Pennsylvania State University
University Park, PA, USA
yml5563@psu.edu

Ting-Hao 'Kenneth' Huang
The Pennsylvania State University
University Park, PA, USA
txh710@psu.edu

ABSTRACT

The proliferation of automated conversational systems such as chatbots, spoken-dialogue systems, and smart speakers, has significantly impacted modern digital life. However, these systems are primarily designed to provide answers to well-defined questions rather than to support users in exploring complex, ill-defined questions. In this paper, we aim to push the boundaries of conversational systems by examining the types of nebulous, open-ended questions that can best be answered through conversation. We first sampled 500 questions from one million open-ended requests posted on AskReddit, and then recruited online crowd workers to answer eight inquiries about these questions. We also performed open coding to categorize the questions into 27 different domains. We found that the issues people believe require conversation to resolve satisfactorily are highly social and personal. Our work provides insights into how future research could be geared to align with users' needs.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**; *Auditory feedback*; *Text input*; **HCI theory, concepts and models**; *Empirical studies in interaction design.*

KEYWORDS

Conversational Systems, Question Answering, Reddit

ACM Reference Format:
Shih-Hong Huang, Chieh-Yang Huang, Ya-Fang Lin, and Ting-Hao 'Kenneth' Huang. 2023. What Types of Questions Require Conversation to Answer? A Case Study of AskReddit Questions. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3544549.3585600

1 INTRODUCTION

Automated conversational systems such as chatbots, spoken-dialogue systems, and smart speakers have become routine in modern digital life. With recent advances in deep learning, today's cutting-edge conversational systems can produce fluent responses to users' messages, find pieces of information as requested, and execute simple voice commands. These systems are designed to quickly deliver concrete answers to well-defined questions. However, the potential of human-to-AI conversations extends beyond this. People have been solving difficult issues by talking to each other for thousands of years. Interaction allows conversational partners to explore ill-defined, complicated problems together. Open-ended discussion allows people to shape their thoughts and stances on complex issues. Unfortunately, the literature has little to say about how conversational systems can be built to support, facilitate, or even participate in such important discussions. Most task-oriented conversational systems have been built with a relatively clear task procedure in mind, *e.g.*, typical user intents, what information is needed to fulfill each intent, steps to take to elicit needed information from the user, and how to accomplish a task. But real-world problems are usually imprecise. The structure, procedure, needed information, and end goals are often unclear or undecided. Everyday questions as common as "What kind of dog should I get?" and "How can I fit into a new environment?" often require back-and-forth discussion to form a helpful answer and can be vastly different for different people. Although chatbots powered by language models such as ChatGPT [14] and YouChat [18] can engage in open-ended conversations to some extent, they are not primarily designed to solve complicated real-world tasks. Instead, they focus on generating human-like responses to various prompts and inputs.

In this paper, we aim to push the boundaries of conversational systems by examining the types of ill-defined, open-ended questions that can best be answered through conversation. We studied the questions people posted on r/AskReddit[1], which tend to be open-ended and loosely defined. AskReddit is an online discussion board (subreddit) of Reddit, a platform on which users can submit open-ended questions to which other users then respond. We extracted one million random questions from the AskReddit subreddit and created a machine-learning classifier to identify questions that asked for help; it identified 129,483. Then we recruited online

CHI EA '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585600

[1] https://www.reddit.com/r/AskReddit/

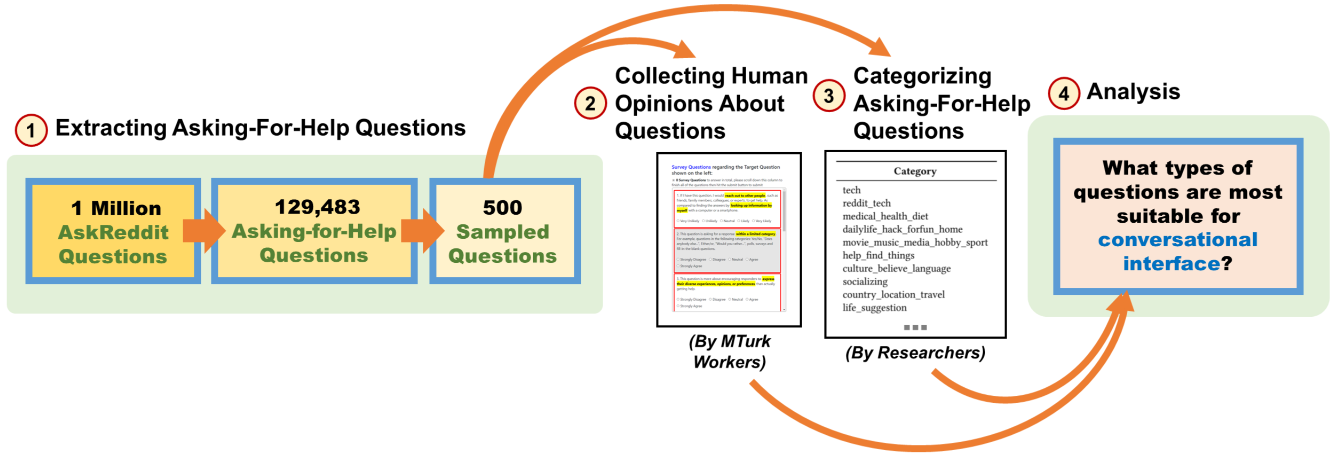

1 Extracting Asking-For-Help Questions
1 Million AskReddit Questions
129,483 Asking-for-Help Questions
500 Sampled Questions
2 Collecting Human Opinions About Questions
(By MTurk Workers)
3 Categorizing Asking-For-Help Questions
Category
tech
reddit_tech
medical_health_diet
dailylife_hack_forfun_home
movie_music_media_hobby_sport
help_find_things
culture_believe_language
socializing
country_location_travel
life_suggestion
(By Researchers)
4 Analysis
What types of questions are most suitable for conversational interface?

Figure 1: The study procedure overview. We (1) sampled 500 asking-for-help questions from one million random AskReddit questions and (2) recruited crowd workers to answer eight inquiries about these questions. Furthermore, we (3) performed open coding on the asking-for-help questions to categorize them, allowing us to (4) gain insight into which sorts of topics require conversation the most.

crowd workers from Amazon Mechanical Turk to answer eight inquiries about 500 randomly sampled asking-for-help questions. These inquiries indicated how much the question required a conversation to be satisfactorily resolved and how the user would most want to get it answered. For example, in one of the inquiries (Q7), we told workers to assume they were asking the question using a computer or smartphone and asked how much they would prefer the answer be provided through a conversation as compared to written formats such as emails. Finally, we performed open coding on the asking-for-help questions to categorize them into 27 different domains, allowing us to analyze which topics MTurk workers believed would require conversation the most. Figure 1 shows the procedure overview of the study.

Instead of asking what a conversational system can do, this work takes a step back and uses a data-driven approach to ask what people *hope* conversational systems can do for them. Our work will inform the development of future systems and help us reflect on the current status of chatbots, spoken dialogue systems, and smart speakers.

2 BACKGROUND

Being able to hold human-like open-domain conversations is one of the biggest challenges in AI. With recent advances in large language models, today’s cutting-edge conversational systems are capable of producing fluent responses to users’ messages [11] and reliably finding requested information [3, 19]. The true value of human conversation lies beyond lightweight chitchat or solving clearly defined tasks such as booking a flight. People talk to each other to navigate complex, ill-defined problems together. Modern intelligent assistants such as Amazon’s Echo promise a future in which conversing with a machine is as easy as talking to a friend. But these conversational systems’ capacity is still far from what talking to a friend can offer. While the latest language models like ChatGPT [14] and YouChat [18] are capable of interacting with the users in a conversational manner, concerns regarding the correctness of the responses provided by such models have been raised, and the limitations of such models are also unclear [4].

Researchers have attempted to bootstrap open-domain conversational systems. A classic example is Evorus, which was initially a human-powered chatbot operated by online crowd workers [7, 9] that automated itself over time [8]. Evorus had crowd workers use a worker interface to propose responses, take notes, and vote to sort others’ replies and identify optimal responses. These collective actions allowed the crowd to converse with the user as a single, consistent conversational partner. More importantly, each action the workers took could be automated over time to gradually move away from human-powered systems. However, one lesson learned from Evorus was that more research is needed to create conversational systems that can solve ill-defined problems [6]. Real-world problems are often complicated and imprecise, and a universally optimal solution may not exist. Supporting or automating such conversations requires approaches beyond taking notes and sorting ideas.

3 METHODS

3.1 Data Preparation

We extracted questions from the `one-million-reddit-questions` dataset [16]. The million questions covered a variety of topics and included questions such as “What is the best story in your family?”, “What frustrates you more than anything?”, “What language/s do you speak?”, and so on. We noticed that the data contained many questions that were meant to engage a large audience on Reddit to elicit responses with diverse viewpoints rather than asking for help. (Table 1 shows some examples of asking-for-help questions.) In this paper, we focused on questions that can benefit from one-on-one conversations with a single conversational partner rather than with a crowd. Therefore, we first built a classifier to extract questions that were actually **asking for help**.

Label	Questions
Asking-for-help	What tasks can only be accomplished by humans, and cannot be accomplished by AI or robots?
Asking-for-help	How to increase reddit trophies and how to get it easily ?
Asking-for-help	If a dog scratches you and doesn't bleed but leaves a mark, will it scar?
Asking-for-help	Trying to get my drivers license after having my permit for 6 months what do i do ?
Asking-for-help	Vietnam war has "fortunate son" as its theme song. What other war has a theme song?
Others	If you were a computer what would your specs be?
Others	What was your favourite period in your life?
Others	If you had to choose a famous person to swap lives with, who would it be?
Others	People of reddit who taught themselves in anything how and why did you do it?
Others	What is life like for you now?

Table 1: Examples of "Asking-for-help" and "Others" categories of questions. Asking-for-help questions are defined as questions people will ask a single agent that have a finite answer.

	Asking-for-help	Others	Total
All	133	1,859	1,992
Train	111	1,483	1,594
Valid	22	376	398

Table 2: Data statistics of the 1,992 annotated Asking-for-help Reddit dataset questions. We split data into train and valid sets using a ratio of 0.8 and 0.2 respectively.

3.1.1 Building a Classifier to Extract Asking-for-Help Questions. To train the classifier to identify questions that were asking for help, one of the authors (A1) annotated 1,992 randomly sampled questions from the dataset of one million Reddit questions (L1). As we are interested in those questions that can be responded to by a conversational agent, questions that people would ask a single ideal agent and have a finite answer were considered valid. Questions that were not considered as asking-for-help questions were those based on personal opinions and experiences of the answerer. Questions intended to generate debate and instigate conflict were also excluded. Inter-coder reliability was investigated by another author (A2) independently annotating 450 randomly sampled questions out of the 1,992 questions (L2) following the coding criteria. The research team also discussed and generated another set of annotations (L3) for the 450 randomly sampled questions that were agreed upon among A1, A2, and another research team member (A3). The inter-coder reliability (Cohen's kappa κ) between each of the annotations (L1, L2, L3) was $\kappa(L1, L2) = 0.545$, $\kappa(L1, L3) = 0.748$, and $\kappa(L2, L3) = 0.802$. Table 1 shows some example sentences; Table 2 shows the statistics indicating data imbalance. The data was split into train/valid sets using the ratio 0.8 and 0.2 respectively. We then fine-tuned DeBERTa [5] (`microsoft/deberta-v2-xxlarge`) using Pytorch [15] and Huggingface [17] for text classification. The hyperparameters used were batch size = 32, learning rate = 1e-5 with the linear scheduler, and warm-up ratio = 0.05. The model was fine-tuned with AdamW optimizer [13] using fp16 precision for 30 epochs.

		Asking-for-help	Others	Macro Avg
	Support	22	376	298
Threshold = 0.5	**Precision**	0.52	0.97	0.75
	Recall	0.55	0.97	0.76
	F1	0.53	0.97	0.75
Threshold = 0.0007	**Precision**	0.35	0.99	0.67
	Recall	0.77	0.91	0.84
	F1	0.48	0.95	0.71

Table 3: Asking-for-help classification performance on the validation set. We searched for a decision threshold in which Asking-for-help recall was higher than 0.7 to encourage the Asking-for-help coverage rate.

Evaluating the Classifier. We evaluated the model every 50 steps and kept the checkpoint with the highest macro f1-score. Table 3 shows the classification performance on the validation set. To avoid unintentionally limiting question types, we adjusted the decision threshold (0.0007) to increase `Asking-for-help` recall to 0.7. The decision threshold was decided by moving the decision threshold from 0.5 to 0 (we moved 5e-5 every step, *e.g.*, 0.5, 0.49995, 0.49990, $\cdots$) and computed `Asking-for-help` recall. We stopped the process and kept the decision threshold once `Asking-for-help` recall reached 0.7 ($\geq$ 0.7). The classification performance using 0.0007 as the threshold is shown in Table 3.

Extracting Asking-for-Help Questions. We applied this text classifier on the entire `one-million-reddit-questions` dataset. It identified 129,483 asking-for-help questions (12.94% of the entire Reddit dataset).

3.2 Collecting Human Opinions About Questions

From the 129,483 asking-for-help questions, 500 questions were randomly sampled for human annotation on Amazon Mechanical Turk (MTurk). Five hundred out of one million questions calculated

#	Aspect	Survey Question
Q1	Reach-Out	If I have this question, I would reach out to other people, such as friends, family members, colleagues, or experts, to get help. As compared to finding the answers by looking up information by myself with a computer or a smartphone. (1) Very Unlikely (2) Unlikely (3) Neutral (4) Likely (5) Very Likely
Q2	Scope	This question is asking for a response within a limited category. For example, questions in the following categories: Yes/No, "Does anybody else...", Either/or, "Would you rather...", polls, surveys and fill-in-the-blank questions. (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
Q3	Eliciting	This question is more about encouraging responders to express their diverse experiences, opinions, or preferences than actually getting help. (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
Q4	Elaboration	This question requires or encourages the responders to further discuss with the asker in order to come up with an appropriate answer. (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
Q5	Duration	Without reaching out to other people, for a layperson with no background knowledge related to this question. How long do you think it would likely take for them to figure out the answer to this question with access to the internet? (1) ≤ 30 minutes (2) 30 minutes-2 hours (3) 2 hours-half a day (4) half a day-1 day (5) ≥ 1 day (6) Undoable
Q6	Conversation	If I have this question, I would prefer to have a conversation regarding the details of the question and have a further discussion with the answerer. As compared to asking the question as is and waiting for the answers. (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
Q7	Format	Suppose I asked this question using a computer or smartphone instead of making phone calls or in-person sessions. In that case, I prefer the answer to be provided through a conversation *e.g.*, via WhatsApp or other messaging applications) compared to other written formats, such as emails, social media replies, or online forums. (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
Q8	Difficulty	Without reaching out to other people to get help, I will be able to answer the question by looking up information by myself with access to a computer or a smartphone. (1) Very Difficult (2) Difficult (3) Neutral (4) Easy (5) Very Easy

Table 4: The eight categories and the inquiries used to collect workers' opinions.

from a 95% confidence level and a 5% margin of error provided a valid sample size for analysis [10]. For each of the 500 questions, we asked nine workers to rate eight aspects using a five-point Likert Scale ranging from (1) Strongly Disagree to (5) Strongly Agree. Table 4 shows the eight aspects we used. Options for Q1 ranged from (1) Very Unlikely to (5) Very Likely. Options for Q5 were (1) 30 minutes or less, (2) 30 minutes-2 hours, (3) 2 hours-half a day, (4) half a day-1 day, (5) 1 day or more, and (6) Undoable. Options for Q8 ranged from (1) Very Difficult to (5) Very Easy.

In the study, each Human Intelligence Task (HIT) contained one asking-for-help question for which each worker was asked to answer the eight survey questions (Table 4). Figure 4 (see Appendix) shows the worker interface. We added a 90-second submission lock on the interface to prevent malicious workers from spamming. The compensation for one HIT assignment was $0.25, which was estimated using an hourly wage of $10. Four built-in MTurk qualifications were also used: Locale (US Only), HIT Approval Rate (≥98%), Number of Approved HITs (≥3000), and Adult Content Qualification.

3.3 Categorizing Asking-For-Help Questions

One author (A1) went through all the 500 sampled questions to get familiar with the data. Open coding was performed to come up with a coding scheme. The process was performed repeatedly until all questions are categorized. A total of 44 mutually exclusive categories were created, and each question belonged to only one category. For simplicity, we merged categories that contain less than five questions into the "Other" category, resulting in a total of 27 categories. Table 5 shows the frequency of the coded categories. Following the coding scheme, another author (A2) independently coded 100 randomly sampled questions from the 500 asking-for-help questions. The inter-coder reliability reached a Cohen's kappa of 0.574.

4 EXPERIMENTAL RESULTS

By comparing the annotated aspects and categories, we formulated three results.

Rank	Category	# Questions	Brief Description
1	tech	67	Technology
2	reddit_tech	47	Reddit-related, Reddit searching, Reddit tech support
3	medical_health_diet	40	Medical, health, or diet
4	dailylife_hack_forfun_home	36	Daily life, home, life hack, for fun or “food for thought” discussion
5	movie_music_media_hobby_sport	31	Movies, music, media, hobby, sport
6	help_find_things	28	Help find things or information by providing description
7	culture_believe_language	20	Culture, beliefs, language
8	socializing	17	Socializing
9	country_location_travel	17	Country, location, traveling
10	life_suggestion	16	General life suggestions
11	science	13	Science
12	history_old_days_future	12	History and discussion about the past or future
13	career	11	Career
14	food	10	Food
15	human_body	10	Human body mechanism and functions
16	legal_regulation	10	Law, legal questions, general regulations
17	news_events	9	News and real-life events
18	social_etiquette	9	Social etiquette
19	mental_health	8	Mental health
20	learning_skills	7	Learning and acquire skills
21	personal_finance	7	Personal finance
22	worldwide_society_effect_impact	7	Worldwide scale discussion/societal changes and impact
23	NSFW_sensitive	6	Not safe for work; not suitable for work and sensitive topics
24	politics	6	Politics
25	relationship	6	Couple relationships
26	religion	5	Religion
-	Others	45	Categories that appear fewer than five times

Table 5: The frequency of the coded categories. Categories with less than five questions are merged into the “Other” category, resulting in a total of 27 categories.

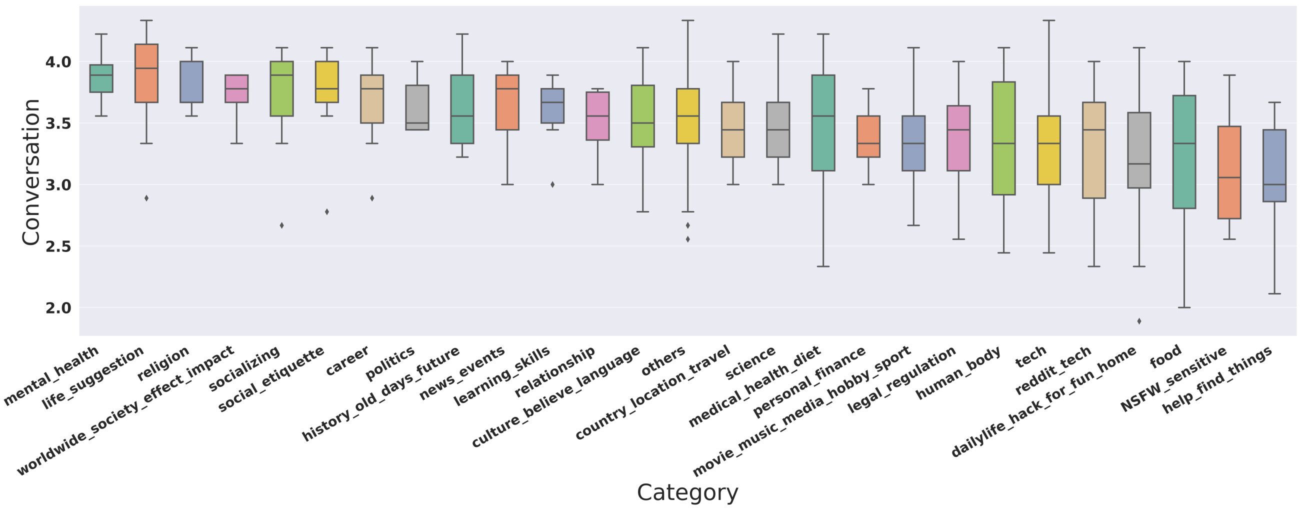


Figure 2: Conversation score (Q6) distribution over different categories. People like to have conversations to consult on questions that do not have clear answers (*e.g.*, mental_health and life_suggestion). For questions that have clearer answers (*e.g.*, help_find_things and tech), conversation is less needed.

Category	Question
tech	Recommendation for a vacuum cleaner?
life_suggestion	What to do when you are feeling lost in life?
work_place	What is the best way to subtly and consistently annoy your coworkers, without them ever realising it's your fault?
others	What happens after we die ?
mental_health	What do you do when you can't get an anxiety-inducing thought out of your head?
medical_health_diet	What happens when you chew a poisonous flower for a few seconds but spit it out?
medical_health_diet	how much pain did you feel after wisdom tooth removal?
medical_health_diet	How to reduce one sided cheek fat?
life_suggestion	How to control emotions? How do people control their emotions when they lost their loved one?
life_suggestion	What steps should I take towards moving out of my parents house? I'm at the ripe old age of 16 when the state of Pennsylvania graciously gives me the chance to operate a motor vehicle. What can I do to get myself headed in the direction of living on my own?
life_suggestion	Is it possible to make a good situation out of any bad situation?
science	What is a fine tuned universe? Why is gravity fine-tuned?
history_old_days_future	Is politics more entertaining now than it was in decades prior?
mental_health	Let me start off by saying, yes I've tried most of the normal avenues, and yet my mind is still filled with thoughts of nihilism. Every moment of my life feels like I'm just waiting. Not for anything in particular, just something. Is there anywhere for people like me to go, and just disappear?
others	This housing market is wild. Is it going to last the next 4 years?

Table 6: Questions with the highest Conversation (Q6) score (≥ 4.22).

4.1 What types of questions are a better fit for conversational UI?

Figure 2 shows the box chart of the Conversation scores over different categories. The categories were sorted descending (from left to right) using the mean Conversation scores. We found that people believe conversations were needed most when questions did not have clear answers, *e.g.*, mental_health, life_suggestion, religion, worldwide_society_effect_impact, socializing, and social_etiquette. Questions that might have concrete responses did not need to be resolved through conversations, *e.g.*, help_find_things, NSFW_sensitive, food, dailylife_hack_for_fun_home, and reddit_tech. See Table 6 for questions with the highest Conversation score.

4.2 Correlation among aspects

To see the relationships among different aspects, we computed the Pearson correlation between all the aspects. Table 7 shows the correlation. We found that Conversation is highly correlated with Eliciting (0.663), Elaboration (0.720), and Format (0.665), suggesting that when a question required a conversation to satisfactorily explore, people believe this question to (*i*) be more related to personal opinions and experiences and (*ii*) require more discussion. Also, in such cases people generally prefer to have a conversation on messaging applications compared to other formats. Since the score for Difficulty is in reverse fashion, (1) being Very Difficult and (5) being Very Easy, the Difficulty score is negatively correlated with most other aspects.

4.3 Category distribution shift

We further compared how the categories were distributed among all the questions and among the questions with high Conversation scores. The category distribution was represented by the percentage of questions within different categories. Questions with high conversation preference (Conversation-desiring) were determined by Conversation score ≥ 3.5. Figure 3 shows the distribution shift. We sorted the categories by the difference between the distribution shifts (*i.e.*, percentage of Conversation-desiring question subtracting percentage in All question) descending from left to right. The figure suggests that the life_suggestion, socializing, and mental_health categories increased more within the Conversation-desiring questions while movie_music_media_hobby_sport, tech, and help_find_things reduced more. This further implied that more personal or social questions are better suited for conversation compared to other types.

5 DISCUSSION

From our results, we identified three areas of discussion.

People want to talk about social situations and personal problems. Our analysis shows that questions people believe require conversation to resolve satisfactorily are **highly social and personal**. Examples include life suggestions, socializing, and mental health. Meanwhile, the questions related to tech or information seeking were considered least requiring of conversation.

	Scope	Eliciting	Elaboration	Duration	Conversation	Format	Difficulty
Reach-Out	0.038	**0.519**	**0.553**	0.191	0.469	**0.502**	−0.035
Scope	-	−0.103	−0.089	−0.104	−0.111	−0.016	0.321
Eliciting	-	-	**0.658**	0.246	**0.663**	**0.607**	−0.186
Elaboration	-	-	-	0.323	**0.720**	**0.672**	−0.326
Duration	-	-	-	-	0.294	0.259	−0.278
Conversation	-	-	-	-	-	**0.665**	−0.323
Format	-	-	-	-	-	-	−0.224

Table 7: Pearson Correlation between different aspects. Bold represents highly correlated (≥ 0.5).

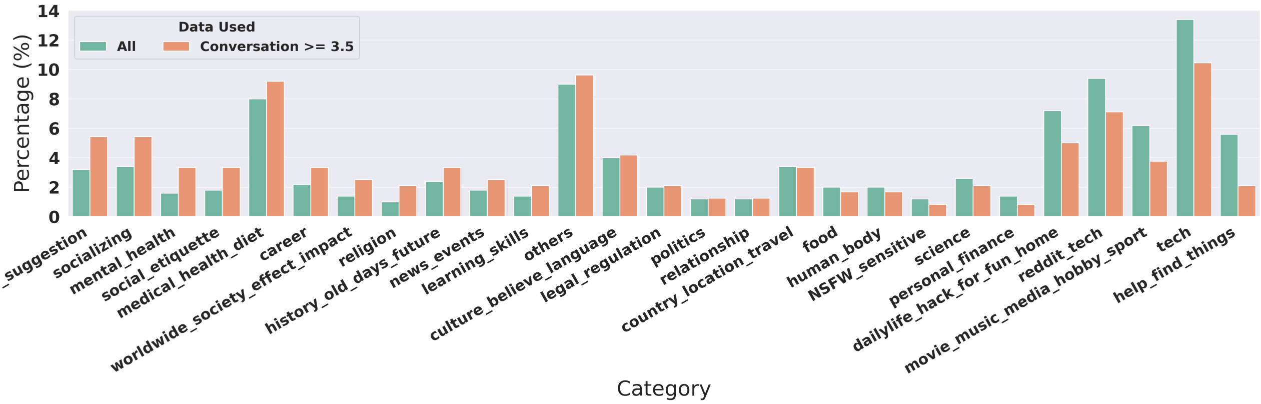


Figure 3: Category distribution shift between all the questions and the ones with higher conversation preference (Conversation score (Q6) ≥ 3.5). Categories with the highest increase in percentage are life, mental health, and socializing.

These findings prompt us to rethink the notion of conversation, especially the differences between producing answers in fluent, natural language and exploring a topic in back-and-forth interaction. Under the broader umbrella of conversational systems, many techniques were created and evolved to achieve the latter, but our findings suggested that in some cases, users might only need the former. Our second conclusion is that the possibility of enabling multiple response channels for conversational systems could be further pursued. For the questions that tend not to require an interactive conversation to resolve, the system can take extra time and resources to prepare the answer and respond via an alternative channel such as email or text messages. This will introduce a new set of technical and UX questions, including how to automatically choose the response channel, customize the user's preference, collect needed information, or ask follow-up questions via multiple channels. Finally, we are aware that a significant body of work has explored using chatbots or conversational agents to provide therapy or mental health support [2, 12]. Even though these questions are often much harder to solve, our results suggest that these attempts are valuable to users.

Some questions require extra attention. Some topics are sensitive, controversial, or potentially harmful. Categories such as politics, religion, NSFW_sensitive, and suicide-related likely need to be handled with extra caution. Our study showed that these types of questions are not rare. Out of 500 asking-for-help questions, six were about politics, five were about religion, six were categorized as NSFW, and one was related to suicide.

Limitations. We are aware of some limitations of this work. First, the automatic classifiers' performance was not perfect. Although we tuned the classifier's parameter to yield high recall, some asking-for-help questions may have been excluded from our study. Second, the scale of our system was relatively small. We could only afford to manually annotate and categorize 500 (each question having nine responses) of the millions of questions posted to AskReddit. Finally, the selection of platforms inevitably imposed biases. The asking-for-help questions were sampled from Reddit, whose users tend to be younger, US-centric, and primarily male [1]. The AskReddit platform also has community norms that encourage questions that generate discussions rather than asking-for-help questions. Using MTurk introduced similar biases.

6 CONCLUSION AND FUTURE WORK

This paper studies what types of questions are most suitable for conversational modality. We recruited online crowd workers to answer eight inquiries about 500 questions posted on AskReddit and performed an in-depth analysis. We found that the questions people believe require conversation to resolve satisfactorily are highly social and personal. Examples include life suggestions, socializing, and mental health. Meanwhile, the questions related to tech or information seeking were considered least requiring of conversation.

In the future, we will develop computational models that automatically recommend the appropriate delivery modality for questions. Such a model would allow intelligent question-answering systems to personalize the communication channel to users.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their constructive feedback, and to the MTurk workers for their participation in our study.

REFERENCES

[1] Pew Research Center. 2022. Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

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

[3] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176* (2022).

[4] Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. "I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. https://doi.org/10.48550/ARXIV.2212.05856

[5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[6] Ting-Hao K. Huang. 2019. On Automating Conversations. *Artificial Intelligence and Work: AAAI 2019 Fall Symposium* (2019).

[7] Ting-Hao K. Huang and Jeffrey P. Bigham. 2017. A 10-Month-Long Deployment Study of On-Demand Recruiting for Low-Latency Crowdsourcing. In *In Proceedings of The fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2017)*. AAAI, AAAI.

[8] Ting-Hao K. Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 10 pages.

[9] Ting-Hao K. Huang, Walter S. Lasecki, Amos Azaria, and Jeffrey P. Bigham. 2016. "Is there anything else I can help you with?": Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Proceedings of AAAI Conference on Human Computation and Crowdsourcing 2016 (HCOMP 2016)*. AAAI.

[10] Glenn D Israel. 1992. Determining sample size. (1992).

[11] Apoorv Kulshreshtha, Daniel De Freitas Adiwardana, David Richard So, Gaurav Nemade, Jamie Hall, Noah Fiedel, Quoc V. Le, Romal Thoppilan, Thang Luong, Yifeng Lu, and Zi Yang. 2020. Towards a Human-like Open-Domain Chatbot. In *arXiv*.

[12] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 31 (may 2020), 27 pages. https://doi.org/10.1145/3392836

[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[14] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[16] SocialGrep. 2021. one-million-reddit-questions. https://huggingface.co/datasets/SocialGrep/one-million-reddit-questions.

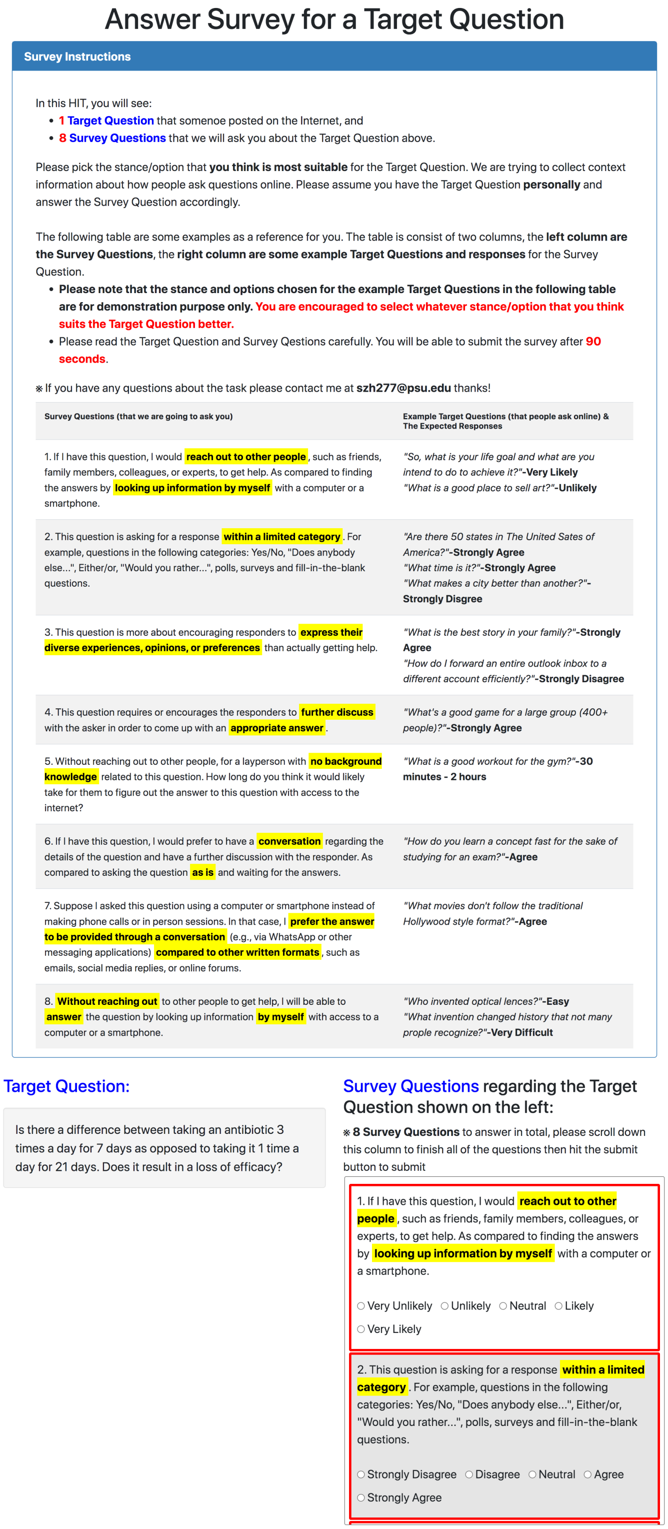
[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[18] You.com. 2022. Introducing YouChat - The AI Search Assistant that Lives in Your Search Engine. https://blog.you.com/introducing-youchat-the-ai-search-assistant-that-lives-in-your-search-engine-eff7badcd655.

[19] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

A SUPPLEMENTARY MATERIAL

Figure 4 shows the MTurk interface for collecting online crowd workers' opinions through the eight questions (Table 4).



Answer Survey for a Target Question

Survey Instructions

In this HIT, you will see:

- 1 Target Question that somenoe posted on the Internet, and
- 8 Survey Questions that we will ask you about the Target Question above.

Please pick the stance/option that **you think is most suitable** for the Target Question. We are trying to collect context information about how people ask questions online. Please assume you have the Target Question **personally** and answer the Survey Question accordingly.

The following table are some examples as a reference for you. The table is consist of two columns, the **left column are the Survey Questions**, the **right column are some example Target Questions and responses** for the Survey Question.

- **Please note that the stance and options chosen for the example Target Questions in the following table are for demonstration purpose only. You are encouraged to select whatever stance/option that you think suits the Target Question better.**
- Please read the Target Question and Survey Qestions carefully. You will be able to submit the survey after **90 seconds**.

※ If you have any questions about the task please contact me at **szh277@psu.edu** thanks!

Survey Questions (that we are going to ask you)	Example Target Questions (that people ask online) & The Expected Responses
1. If I have this question, I would **reach out to other people**, such as friends, family members, colleagues, or experts, to get help. As compared to finding the answers by **looking up information by myself** with a computer or a smartphone.	*"So, what is your life goal and what are you intend to do to achieve it?"***-Very Likely** *"What is a good place to sell art?"***-Unlikely**
2. This question is asking for a response **within a limited category**. For example, questions in the following categories: Yes/No, "Does anybody else...", Either/or, "Would you rather...", polls, surveys and fill-in-the-blank questions.	*"Are there 50 states in The United Sates of America?"***-Strongly Agree** *"What time is it?"***-Strongly Agree** *"What makes a city better than another?"***-Strongly Disgree**
3. This question is more about encouraging responders to **express their diverse experiences, opinions, or preferences** than actually getting help.	*"What is the best story in your family?"***-Strongly Agree** *"How do I forward an entire outlook inbox to a different account efficiently?"***-Strongly Disagree**
4. This question requires or encourages the responders to **further discuss** with the asker in order to come up with an **appropriate answer**.	*"What's a good game for a large group (400+ people)?"***-Strongly Agree**
5. Without reaching out to other people, for a layperson with **no background knowledge** related to this question. How long do you think it would likely take for them to figure out the answer to this question with access to the internet?	*"What is a good workout for the gym?"***-30 minutes - 2 hours**
6. If I have this question, I would prefer to have a **conversation** regarding the details of the question and have a further discussion with the responder. As compared to asking the question **as is** and waiting for the answers.	*"How do you learn a concept fast for the sake of studying for an exam?"***-Agree**
7. Suppose I asked this question using a computer or smartphone instead of making phone calls or in person sessions. In that case, I **prefer the answer to be provided through a conversation** (e.g., via WhatsApp or other messaging applications) **compared to other written formats**, such as emails, social media replies, or online forums.	*"What movies don't follow the traditional Hollywood style format?"***-Agree**
8. **Without reaching out** to other people to get help, I will be able to **answer** the question by looking up information **by myself** with access to a computer or a smartphone.	*"Who invented optical lences?"***-Easy** *"What invention changed history that not many prople recognize?"***-Very Difficult**

Target Question:

Is there a difference between taking an antibiotic 3 times a day for 7 days as opposed to taking it 1 time a day for 21 days. Does it result in a loss of efficacy?

Survey Questions regarding the Target Question shown on the left:

※ **8 Survey Questions** to answer in total, please scroll down this column to finish all of the questions then hit the submit button to submit

1. If I have this question, I would **reach out to other people**, such as friends, family members, colleagues, or experts, to get help. As compared to finding the answers by **looking up information by myself** with a computer or a smartphone.

○ Very Unlikely ○ Unlikely ○ Neutral ○ Likely ○ Very Likely

2. This question is asking for a response **within a limited category**. For example, questions in the following categories: Yes/No, "Does anybody else...", Either/or, "Would you rather...", polls, surveys and fill-in-the-blank questions.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Figure 4: Interface for MTurk workers.