CoArgue: Fostering Lurkers' Contribution to Collective Arguments in Community-based QA Platforms

Chengzhong Liu
chengzhong.liu@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Shixu Zhou
szhouav@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Dingdong Liu
dliuak@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Junze Li
jlijj@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Zeyu Huang
zhuangbi@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Xiaojuan Ma*
mxj@cse.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

ABSTRACT

In Community-Based Question Answering (CQA) platforms, people can participate in discussions about non-factoid topics by marking their stances, providing premises, or arguing for the opinions they support, which forms "collective arguments". The sustainable development of collective arguments relies on a big contributor base, yet most of the frequent CQA users are lurkers who seldom speak out. With a formative study, we identified detailed obstacles preventing lurkers from contributing to collective arguments. We consequently designed a processing pipeline for extracting and summarizing augmentative elements from question threads. Based on this we built CoArgue, a tool with navigation and chatbot features to support CQA lurkers' motivation and ability in making contributions. Through a within-subject study (N=24), we found that, compared to a Quora-like baseline, participants perceived CoArgue as significantly more useful in enhancing their motivation and ability to join collective arguments and found the experience to be more engaging and productive.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Collective Arguments, CQA Platforms, Lurker Support

ACM Reference Format:

Chengzhong Liu, Shixu Zhou, Dingdong Liu, Junze Li, Zeyu Huang, and Xiaojuan Ma. 2023. CoArgue: Fostering Lurkers' Contribution to Collective Arguments in Community-based QA Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3580932

*Corresponding author

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580932

1 INTRODUCTION

Mainstream Community-Based Question Answering (CQA) platforms such as *Quora*[1] and *Yahoo! Answers*[2], powered by collective intelligence, have attracted millions of users to digest and share knowledge daily [112–114]. On such platforms, people participate in online discussions by answering questions posted in the CQA community and forming question threads on different topics [27, 47, 66, 113]. Particularly within the non-factoid [31, 87] (*a.k.a.,* conversational [38, 113] or dynamic [68]) type question threads, it is a common practice for CQA users to declare their stances and provide premises to argue for the opinions they support around the concerning subjects [5, 59, 126, 137]. Aggregations of such argument-related elements from online discussions centering on specific topics can be viewed as *"collective arguments"* co-created by members of CQA communities [119].

Estimated question threads involving collective arguments account for over half of the User-Generated Content (UGC) contributed by CQA users [68]. Different from the other types of question threads focusing purely on information or resource seeking [68], argumentation-oriented question threads offer people an opportunity to exchange their views and learn from the crowd towards those inconclusive topics without fixed or best answers [31, 87, 113]. For example, the set of argumentative answer posts in a question thread discussing whether one should invest in bitcoins[3] forms a repository of collective arguments concerning this emerging form of investment. While each argumentation tends to have a single standpoint [1, 124, 134], the whole repository provides a relatively comprehensive view of diverse and even opposing stances: some contributors support bitcoin investment for the convenience of cross-border transactions, while others challenge it as risky. Such collective arguments may help break the echo chamber and foster critical thinking [37] of both discussion participators and other

[1] https://www.quora.com
[2] https://answers.yahoo.com/
[3] https://www.quora.com/Should-I-invest-in-Bitcoin

readers by supporting minority voices underrepresented in existing answer posts [137]. Moreover, these question threads also provide social incentives to future contributors to enrich and expand the exchange of opinions [33, 38].

Collective arguments developed by community intelligence on CQA platforms [113] typically feature the interconnection between the arguments. As one may observe in the bitcoin example, some contributors do not specify their stances and only provide premises to support several existing claims in others' answers that belong to different viewpoints [113, 127]. On the one hand, the participation barrier for CQA members is lowered as they only need to identify and fill the gaps of the existing collective arguments they are interested in instead of creating independent and standalone arguments. In this process, they can also sharpen their argumentative writing skills [59] and potentially meet their self-fulfillment if their writings receive acknowledgment (*i.e.,* upvotes) from others [33, 91, 101]. On the other hand, the content generated by individual users can be limited [71, 120, 139], CQA platforms need to maintain a relatively large contributor base to ensure the sustainable development of such collective arguments [7, 87, 111]. However, most CQA users are lurkers who frequently visit CQA sites but rarely publish any content [60, 71, 112–114]. Based on the behavior change theories [39, 116], CQA platforms would need to sufficiently boost the motivation and ability of lurkers to turn them into active contributors. Still, the existing methods employed by them are not that effective. To be more specific, currently, the most direct way to motivate a user to answer a question post is by invitation from other CQA members, which works mainly for active users with more robust social connections in the community; lurkers, however, seldom receive such invitations with their limited exposure to the community [109, 113, 136]. There is even less support by CQA platforms concerning users' ability to digest and add new, meaningful information to the collective arguments (*e.g.,* claims or premises, dependent or independent from existing contents) [59, 79].

Previous Human-Computer Interaction (HCI) works proposed many solutions to incentivize lurkers of online communities and to enhance argumentative writing ability, respectively. Prior research persuading online community members to participate mainly targeted closely connected or task-specific communities. For instance, Li *et al.* suggested that collective socialization as part of the onboarding practices could motivate newcomers and lurkers to contribute more to the closely connected online communities such as project collaborations [62]. As for task-specific communities, Kaur *et al.* encouraged information exchange between patients and caregivers of online health communities by enhancing sense-of-community with spiritual support such as calm and comfort [46]. While informative, these findings may not directly apply to mainstream CQA communities that are generally loosely connected and involve a broad scope of discussions [43, 113]. Regarding technological support for argumentation, existing tools focus on aiding users in producing standalone writings. For example, Wambsganss *et al.* designed a chatbot tutor that could provide feedback to assist students in developing better arguments [124]. Xia *et al.* implemented an interactive visual system that could improve the persuasive skills of the users [134]. Nevertheless, these tools may not be as effective for facilitating the construction of collective arguments. Many took an example-based approach, showing similar arguments by others as inspirations [5, 113, 119]. Lurkers, however, are likely to be discouraged by such high-quality examples, considering the bar for contribution is too high for them, and there leaves little room for them to participate [56, 118]. In brief, given that collective arguments account for more than half of the CQA contents [68, 113], there is a call for CQA platforms to effectively engage lurkers in contributing to collective arguments via adequate motivation and ability support. Otherwise, the thriving of CQA communities could not be guaranteed and might even shrink without sufficient contributors [7, 87, 111, 113].

To fill this gap, we conducted a Formative Study (N=11) to understand what specific obstacles lurkers of CQA platforms face when contributing to the collective arguments. The findings revealed that the lurkers faced obstacles associated with confidence and willingness from the motivation side. In contrast, from the ability side, they found it challenging to digest existing collective arguments and write good answer posts. Based on the derived Design Goals, we designed and implemented CoArgue [4], an interactive system aiming to reduce the obstacles lurkers faced when contributing to collective arguments of CQA platforms. Founded on a Natural language processing (NLP) pipeline that could extract and organize argumentative information, CoArgue fostered lurkers' contributions to collective arguments by **1)** highlighting claims and premises in answer posts, **2)** navigating users through the claims visually, **3)** guiding users to develop their answer posts with a chatbot, and **4)** encouraging users with additional feedback.

Finally, we conducted a within-subject study to evaluate CoArgue's efficacy in providing motivation and ability support, supporting engaging interaction experience, boosting output quality, and maintaining system usability with a Quora-like interface as the Baseline. We recruited 24 CQA lurkers to join two separate and counterbalanced answer-post writing sessions on collective arguments with topics in which they had at least moderate interest and knowledge. Results suggested that participants perceived CoArgue as significantly more helpful in enhancing their motivation and ability to join collective arguments. They also found the interaction experience to be more engaging and composed answer posts with more argumentative elements, *i.e.,* claims and premises. Moreover, according to the participants, although the usability was reduced with additional interaction designs, they were more willing to use CoArgue again compared to the Baseline. We further summarized critical design implications for future tools like CoArgue.

The key contributions of this work are threefold: **1)** CoArgue, an interactive visualization system that fostered lurkers' contributions to CQA collective arguments, **2)** A within-subject study demonstrated the usefulness and effectiveness of CoArgue compared to a Quora-like Baseline system, and **3)** Design considerations that could guide future designs on how to encourage and support CQA lurkers to join collective arguments.

2 RELATED WORK

In this section, we surveyed literature on the connection between lurkers and online communities, visualization and NLP methods to process arguments, and existing techniques to support CQA users.

[4] Open-sourced at https://github.com/Zascc/CoArgue-CHI2023

2.1 The Connection Between Lurkers and Online Communities

In the Introduction, we listed a few practices on de-lurking for CQA and online communities [46, 59, 62, 79, 109, 113, 136], and in this section, we further surveyed previous works to understand the lurking behaviors in online communities. Chen *et al.* found that in an online collective learning community, lurkers who wrote few individual posts could make a comparable number of contributions to others in group assignments, which implied that lurkers could be motivated by the community members they connected to [13]. Lampe *et al.* identified that the sense of belonging, which can make users feel confident and comfortable in online communities [97], is one of the key elements for lurkers to start making continuous contributions to online user-driven encyclopedias [58]. Kim *et al.* revealed that mobilizing lurkers, even only with a chatbot, could increase their participation in an institution's online community [51]. Moreover, existing research also regarded lurking as a necessary process for users to get familiar with the communities to develop sufficient confidence to contribute [81, 85].

In summary, the connection between communities and participants is decisive in the level of user participation. Participants in loosely connected communities like CQA perceive less presence of other members and therefore tend to behave more like lurkers [43, 113]. In this work, we aimed to start by strengthening the connection between CQA lurkers and the collective arguments, designing a tool to facilitate answer post composing.

2.2 Visualization and NLP Techniques to Process Arguments

Arguments generally consist of claims & premises, and the relationships between these components are complex [125, 134]. Therefore, researchers designed visualization tools with diverse structures and layouts to adapt to particular argument styles [49, 50]. Khartabil *et al.* summarized existing argument visualization tools into four layout categories [50] and later proposed a combined visualization method, including tree visualizations, content display, and interactive navigation [49]. Wambsganss *et al.* designed a learning support system to improve students' argumentation skills by labeling the argumentative components & relations and providing the evaluation scores [125]. Xia *et al.* built the Persua system to visualize the relationship between claims and different types of premises to enhance the persuasiveness of arguments [134]. However, most works focused on per-argument visualization, not considering connections between posts which are essential for forums-based communities like CQA platforms [64]. To fill the gap, we designed CoArgue that not only presented arguments in a systematic claim-premise structure, but also organized their argumentative components in a holistic manner.

Apart from the visualization methods, processing arguments with NLP techniques have been developing rapidly in recent years. Goudas *et al.* proposed a two-step method based on SVM [16] and CRF [55] to extract arguments from social media [35]. Akiki *et al.* implemented an argument extraction pipeline with BERT [20] and GPT-2 [96], achieving the best retrieval performance in Touché shared task [6]. Other models, such as BART [61] and PEGASUS [138], provided state-of-the-art performances in text summarization tasks similar to argument extraction. Besides, Dumani *et al.* proposed a framework for argument retrieval systems that could facilitate argument understanding. [23]. Following their paths, we constructed a pipeline that could firstly identify arguments with related claims and premises, then cluster similar ones, and finally respond to users' query with related argumentative components.

2.3 Technologies to Support CQA Users

Developing tools to support users of CQA platforms has always been a hot research topic. One mainstream is to provide text summarization and visualization of the topic and central idea on a single post [18], of the entire thread [128], or by each author [41, 89]. Such summarization and visualization tools are beneficial for information digestion but cannot motivate lurkers to participate in discussion and contribute to the collective arguments. From another perspective, Hoque *et al.* developed CQAVis to visualize information usefulness, relevance, and richness in an entire CQA thread [42]. Although the visualizations could manifest current coverage of collective arguments, thus potentially encourage users to write, it only ranked and filtered answer posts according to generic metrics like *similarity* and *usefulness* [42], without supporting users to dissect the detailed content in each post. Taking inspiration from the aforementioned works, we designed CoArgue with two purposes: **1)** motivating and engaging users with interactive features and **2)** assisting users in digesting CQA collective arguments effectively via proper text summarization and visualization.

3 FORMATIVE STUDY

To more comprehensively understand the causes of the lurking behaviors on CQA platforms, we conducted semi-structured interviews with 11 CQA lurkers. The study revealed nine obstacles that hindered them from contributing to collective arguments, which further guided the design of CoArgue, a computer-aided system that supports lurkers to overcome these obstacles.

3.1 Participants and Procedure

With the approval of the institution's IRB, we recruited 11 participants for the formative study (four female, six male, 1 prefer not to say; age range 20-29; indexed PF1 to PF11) through online advertising, social media, and word-of-mouth at a local university. All participants are self-reported lurkers of mainstream CQA platforms. Although they frequently visit the platforms (two participants browse CQA posts every day, six 4-6 days a week, and three at least once a week) and often experience impulse to join CQA discussions, they make few contributions (all composing no more than three answers posts over the past year). We conducted a semi-structured interview with each of the participants to understand their lurking behaviors in CQA communities and why they seldom contributed to the collective arguments. More specifically, we asked what topics of the collective arguments (related question-answer threads) usually attracted their interest, why they were "so silent" towards those topics, what were the major factors preventing them from writing responses, and what kinds of technical support could make them more active contributors to those collective arguments. In addition, we also learned about their experience

Table 1: Obstacles reported by lurkers when contributing to collective arguments.

Category	Sub-category	Obstacles
Motivation	Confidence	**O1**. Consider the contribution from oneself is trivial
		O2. Assume the writing has to be comprehensive
	Willingness	**O3**. Neglect the importance of their contributions
		O4. Worry that contributing would be time-consuming
Ability	Digesting	**O5**. Encounter difficulty on digesting unstructured arguments
		O6. Feel overwhelmed by too many arguments
	Writing	**O7**. Suppose the angle to contribute is hard to locate
		O8. Find it challenging to organize the scattered insights
		O9. Need to maintain the engagement during the writing

with *collective arguments*: the structure of collective arguments in their understanding, current strategies to digest them, and potential approaches to join them. The interview typically lasted around 30-40 minutes, and each of the participants received a $10 gift card.

Following the procedures of literature [3, 52, 67, 94], the interviews were audio-recorded and transcribed to text, and two researchers independently conducted open coding over eight transcripts to get the initial codes. Then the whole research team derived the categories/subcategories and constructed a codebook accordingly over group discussions. Next, the team reread the eight transcripts and applied the codebook. Finally, the team conducted three more interviews and used the codebook to code their transcripts. As no new codes emerged (*i.e.,* reaching saturation [106]), they stopped to carry out more interviews.

3.2 Findings

In general, the participants mentioned a wide range of topics on which they were interested in the communities' opinions, including abortion rights, autopilot cars, and bitcoin investment, to name a few. Then, through the inductive thematic analysis [8], we identified nine obstacles that prevent the participants from contributing to the collective arguments on these topics, as noted by more than half of them. We further categorized these obstacles into *motivation* and *ability* [39, 116], as detailed in Table 1.

3.2.1 Motivation. Most participants (9/11) reflected on the obstacles to their motivation to become a contributor to CQA collective arguments. We further divided the points into confidence- and willingness-related issues. Many participants seemed to lack confidence in the value they could generate, as they felt that the existing answers were perfect enough (**O1**, seven people) or believed that their argumentative answer posts had to be comprehensive enough to be useful (**O2**, eight people). Specifically, PF10 addressed that *"high voted posts are somewhat intimidating"*, thus discouraging him from posting as he believed that he could not write anything comparable to those top answers. For PF5, she even feared that her posts might bring *"negative effects to the discussion [the development of the collective arguments]"* as her personal experience to share might be biased and misleading.

A related obstacle was also located on the willingness side: participants generally neglected the importance of their contributions to the collective arguments (**O3**, six people), probably due to the loosely coupled nature of the CQA communities [113]. *"Unlike those team works where I can immediately feel the importance of my role, feedbacks after making contributions to CQA are very limited . . . I need to wait for a long time to get some upvotes."* PF3 said. He also made similar speculations as the previous existing research [113], sharing the feeling that *"the [perceived] social distance between the answerers of the existing posts and me is simply too far away!"*. Participants also worried that coming up with a good answer would be time-consuming (**O4**, eight people). PF1 commented that he usually read the feeds on the CQA platforms for entertainment and thus was reluctant to spare more time to make contributions.

3.2.2 Ability. All participants encountered barriers to their ability to add building blocks in the collective arguments on CQA platforms, which could be further divided into digestion- and writing-related ones. Many participants usually found it challenging to process the unstructured arguments made by others (**O5**, seven people), while others were overwhelmed by the quantity of them (**O6**, seven people). Especially three participants (PF4, PF6, PF7) said these issues discouraged them from digesting the elements of the arguments comprehensively; instead, they only searched for the keywords of interest without thoroughly exploring more, which further hindered them from posting their answers.

As for the writing process, the participants considered it hard to find a proper angle to join the collective arguments (**O7**, nine people) and organize their scattered thoughts into text (**O8**, six people). For example, PF2 said that he would feel good if he could enrich repositories of existing answers, but in most cases, he did not know *"what novel stuffs to contribute"*. PF8 reported that she usually got some random thoughts when reading the existing answers, and it would be good to *"have somewhere to drop them down and eventually transform the points to outline for writing"*. Apart from the obstacles to content composition, seven participants also reported the need to maintain engagement during the writing process (**O9**).

3.3 Design Goals

Based on the Formative Study findings and existing research, we derived the following design goals (**DG**s) for a computer-aided system to effectively support lurkers to contribute to the collective arguments on CQA platforms.

DG1 Encourage users to participate in collective arguments and communicate the envisioned value of their contributions. Lurkers of online communities typically underestimate the value of their participation [118]. In addition to this,

the participants in our formative study generally pictured a high bar for an argumentative answer post to be acceptable (**O1**, **O2**). *"I felt it is my responsibility to write something that is 100% beneficial to others without any potential side effects"*, said PF11. Consequently, following the practices of existing work empowered by the sense-of-community theory [46, 76, 100], it would be helpful to imply to lurkers that developing collective arguments is a community-wide effort, and there is no need for a single individual to do it all. In conclusion, informing lurkers of the potential value of their intended contribution angle, extensive or not, could be useful for strengthening their willingness and writing capabilities to make the post (**O3**, **O7**).

DG2 Use technology to support an engaging and efficient interaction experience with the collective arguments. Human engagement is one of the key elements to design computer-aided systems aiming to help users generate better quality output [22, 44, 69, 83]. Participants in our formative study also desired an engaging while efficient contribution process to reduce their perceived workload when joining the collective arguments (**O4**, **O9**). PF5 added that, *"if the composing [process] feels like entertainment, I would definitely output more"*.

DG3 Present a holistic view of the collective arguments and facilitate the digestion process. Collective arguments are repositories of argumentative elements from user-generated contents on the CQA platforms, typically unstructured and large in volume [88]. Existing study [105] suggested that the challenge of digesting such information could be alleviated if the data source could be efficiently navigated by users (**O5**, **O6**). Moreover, considering that a summative representation of compounded contents can ease people's understanding process [25], users could be more focused on their contributions with reduced cognitive load (**O9**).

DG4 Help locate what to contribute based on the existing contents of collective arguments. As mentioned by participants, pinpointing what to contribute and framing the answers are non-trivial obstacles for them to join collective arguments (**O7**, **O8**). Particularly, PF4 expressed interests in supporting minority voices she agreed with but had trouble finding what had been explored among the large pool of mainstream comments.

4 COARGUE SYSTEM

According to the Formative Study participants, a single argumentative answer post of collective arguments typically has one overall stance associated with several specific claims, but they may not be supported by premises. Therefore, within a repository of collective arguments, CQA users would face an overwhelming amount of diverse yet similar claims supported by an arbitrary number of premises. In light of the related literature (section 2.2), we developed the taxonomies below to facilitate the design of CoArgue:

- **Stance**: people's sentiment on a particular topic (*e.g.,* positive, neutral, and negative) [2].
- **Claim**: the proposition to convey one's attitude or stance on a particular topic [40].

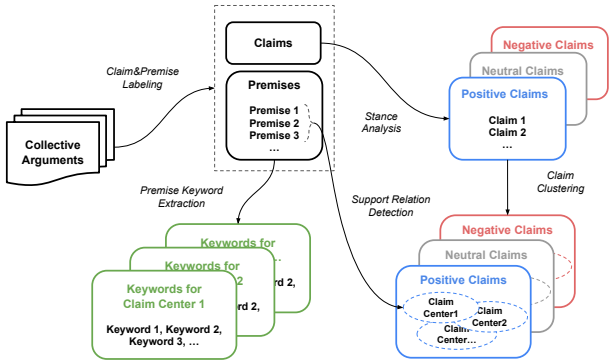

Figure 1: The NLP pipeline of CoArgue.

Table 2: The information load of the selected collective arguments for evaluation.

Topic	# Claims	# Claim Centers	# Premises
Bitcoin Investment	75	14	147
Self-driving Car	82	12	147

- **Claim Center**: the cluster of claims from collective arguments (section 4.1.2).
- **Premise**: the evidence or justification to support the corresponding claims [40].

4.1 Collective Arguments Processing

As suggested by Formative Study participants, we selected and crawled two repositories of collective arguments from *Quora*, one of the mainstream CQA platforms [42, 114], as a starting point to construct CoArgue:

- **Bitcoin Investment**: "Should I invest in Bitcoin?[5]"
- **Self-driving Car**: "Would you get into a self-driving car?[6]"

To control the variables for fair evaluation (section 5.1), two repositories had balanced information load (Table 2); both were related to high-tech industries widely discussed recently [32, 48]. Figure 1 demonstrates an overview of the NLP pipeline.

4.1.1 Claim and Premise Labeling. Claims and premises are considered two essential components in an argumentative text [40, 115]. To automatically extract the claims and premises from the collective arguments, we trained a BART [61] model based on the *Change My View* dataset[7] with corresponding human labels [40]. The dataset contains rich CQA type of online discussion data and reliable human labels [40] and has been used for the HCI research related to argumentative writings [12, 134].

We performed a 7:3 split for training and testing (417/180 Reddit posts and replies) on the dataset and adopted the pre-trained *sshleifer/distilbart-cnn-12-6* model [131]. As shown in Table 3, the model achieved a high ROUGE-L similarity score [63] of 70.16% and 75.68%, respectively, on the test set compared with the ground

[5]https://www.quora.com/Should-I-invest-in-Bitcoin
[6]https://www.quora.com/Would-you-get-into-a-self-driving-car
[7]https://www.reddit.com/r/changemyview/

Table 3: Performance (in %) of claim and premise labeling algorithm on *Change My View* dataset.

Target	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L-sum
Claim	71.44	67.81	70.16	70.73
Premise	76.94	72.88	75.68	76.57

truth. Then we deployed the trained model to label the claims and premises of the crawled collective arguments.

4.1.2 Claim and Premise Processing. To support the functionalities of **DG3** and **DG4**, we further processed the extracted claims and premises, as detailed in the below paragraphs and Figure 1.

Stance Analysis. We analyzed the author's stance on each claim in terms of positive, neutral, or negative. We fed each claim into the well-acknowledged sentiment analysis model *cardiffnlp/twitter-xlm -roberta-base-sentiment* [4, 131] trained by *HuggingFace*[8], and acquired the sentiment label. Claims with the same stance were grouped for further clustering.

Claim Clustering. For each stance group, we clustered the claims for more structured digestion. We first implemented a pre-trained model of sentence BERT, namely *all-MiniLM-L6-v2* [98] to encode claims into high-dimensional vectors. Then, we performed Principal Component Analysis (PCA) [92] on these vectors to reduce dimension. Compared to other dimension reduction techniques (*e.g.,* MDS [92] and TSNE [95]), PCA maintains the density of original vectors [75] and is efficient enough to be deployed in a real-time system [74]. Finally, applying the Affinity Propagation algorithm [30] we clustered the claims with the centroid of each cluster (*i.e.,* the core claim) as the claim center.

Premise Keyword Extraction. After processing the claims, we extracted keywords from premises to assist users' argument writing in the writing phase (sections 4.2.3 and 4.2.4 below). We collected all the associated premises that supported claims within each claim center. The *support relation* (Figure 1) was determined by the proximity of the premise to the claim in a logical order [124, 134]. Then we concatenated these premises into one paragraph for each claim center and consequently fed them into the pre-trained keyBERT model [36] (with [98] providing sentence embedding) to generate keywords.

4.2 User Interface

Based on the designed NLP pipeline (section 4.1) to process collective arguments, we built the user interface of CoArgue to fulfill the proposed Design Goals.

4.2.1 Answer View. The overall structure of the Answer View was similar to the Quora interface, including each user's name, publish date, number of upvotes as meta-information, and the functionality to collapse or expand the answer post. In each answer post, we highlighted the claims and premises (Figure 2, ①). Claims were highlighted according to their stances (section 4.1.2), while all premises were in the same color. In this way, users could quickly catch the authors' stances, specific claims, and relevant premises (**DG3**).

4.2.2 Navigation View. To help users grasp a holistic understanding of the collective arguments, the Navigation View (Figure 2, ③) listed the claim centers in each stance group (section 4.1.2) and illustrated the overall stance distribution (**DG3**, **DG4**). Each claim center was encoded with its popularity defined as the number of claims it contained to total (**DG4**). Upon clicking, users could explore all the associated claims (**DG3**), and they might further click the claim and jump to the original answer post in the Answer View (Figure 2, ④).

4.2.3 Chatbot View. Apart from the previous two Views, we implemented a Chatbot View (Figure 2, ⑤) as an alternative way to interact with the collective arguments because they can improve the user's confidence [11, 78], provide an engaging interaction experience [51, 93, 102], and facilitate content creation [108, 130] (**DG1**, **DG2**, and **DG4**). We developed the chatbot using the *RASA*[9] framework and further equipped it with self-mockery languages to engage users and improve their confidence [11, 65].

The main interaction logic of the chatbot is shown in Figure 3. At the very beginning, the chatbot would greet the user and invite them to compose the answer post together. It would start by asking users about their stance in mind and provide encouragement if the user currently got no idea. Then the chatbot would guide users to walk through the claim centers from the less popular ones to help them either formulate their own claims or choose from the existing ones. After pinpointing one specific claim, users could opt to discuss with or get hints from the chatbot (powered by keywords extracted, section 4.1.2) before heading to the final writing stage.

At any time of the interaction, users could check the notes generated from the interaction history by the chatbot. They could also click the premise in the Answer View, and the chatbot would provide the claim center it supported as well as other premises supporting this claim center (Figure 2, ② and ⑥).

4.2.4 Writing Interface and Contribution View. Similar to Quora, users could start writing anytime in a dedicated interface. Yet, CoArgue would automatically pre-fill the draft with the stance, claim, and keywords that the user had discussed with the chatbot (Figure 4, ①) such that they could start writing easier (**DG2**, **DG4**).

After users submitted their arguments, the Contribution View (Figure 4, ②) would be shown to inform users of their contributions to collective arguments (**DG1**, **DG2**). Supported by our NLP pipeline (Figure 1), the statistics would include the position of the user's writing in length and their contributions to the stance group in terms of the added claims. Eventually, after users confirmed the submission, they could see their post in the Answer View with highlighted claims and premises. The stacked bar plot in Navigation View would also update to indicted their contributions to a specific stance group (Figure 2, ③).

5 EVALUATION

For future work to reproduce our Results, we presented the detailed setup of our experiment, including the study procedures,

[8]https://huggingface.co/

[9]https://rasa.com/

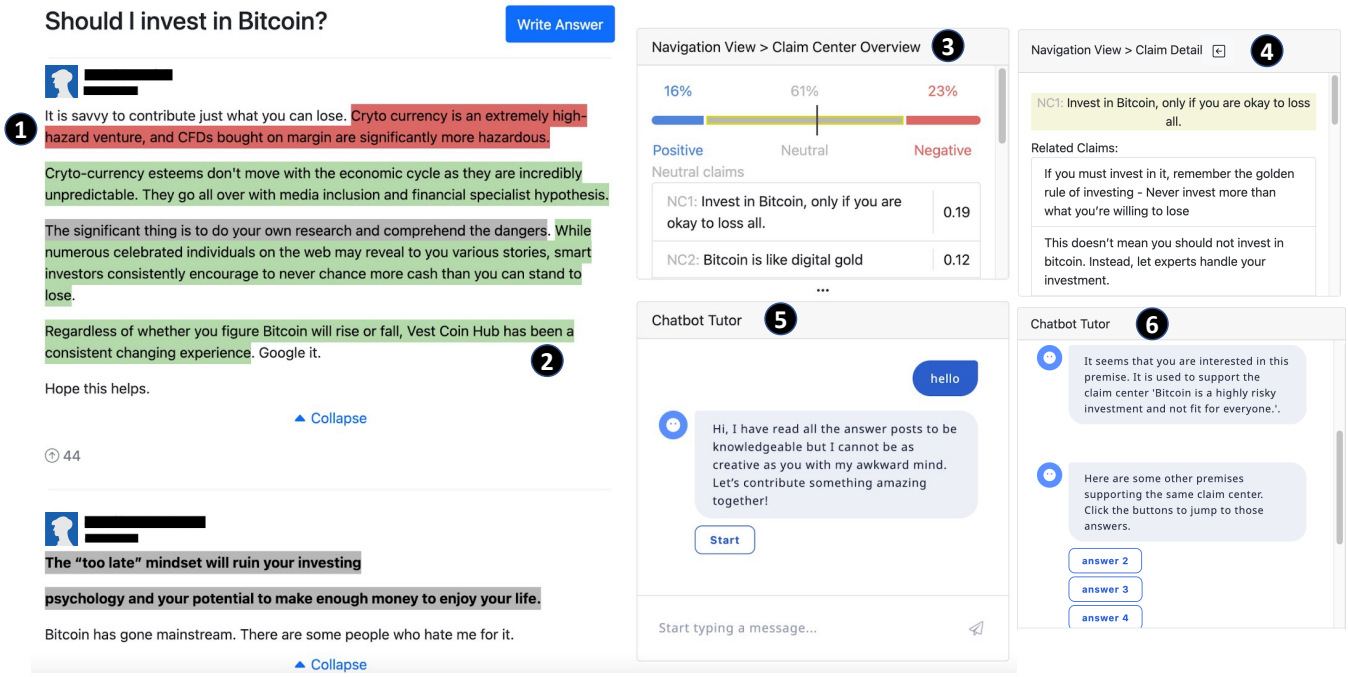

Figure 2: The interface of CoArgue.

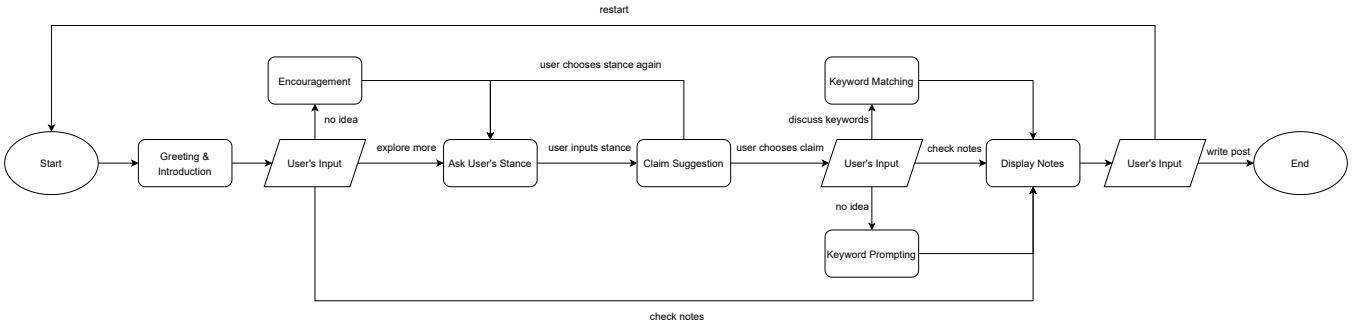

Figure 3: Logic flow of the Chatbot View.

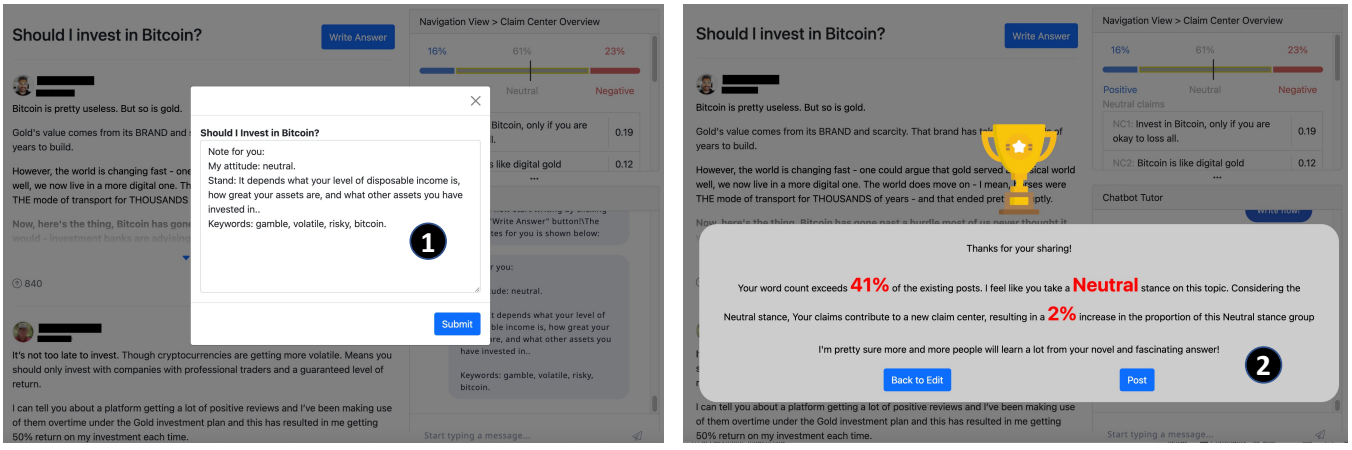

Figure 4: The writing interface and the Contribution View.

Table 4: Demographics of all the participants, including participants' ID, gender, age, their preferred view during the interaction with CoArgue and whether they had qualified writing output (*N.Q.* = Non Qualified).

ID	Gender	Age	Preferred View	Writing Output	
				Baseline	CoArgue
P1	Female	21	Navigation	✓	✓
P2	Male	23	Chatbot	✓	✓
P3	Male	21	Navigation	N.Q.	N.Q.
P4	Female	22	Navigation	✓	✓
P5	Male	26	Navigation	✓	✓
P6	Male	29	Chatbot	✓	✓
P7	Male	23	Navigation	N.Q.	✓
P8	Female	25	Navigation	✓	✓
P9	Male	26	Navigation	✓	N.Q.
P10	Female	22	Chatbot	N.Q.	✓
P11	Female	22	Balanced	✓	✓
P12	Female	25	Chatbot	N.Q.	N.Q.
P13	Male	24	Navigation	✓	✓
P14	Male	24	Navigation	N.Q.	N.Q.
P15	Male	22	Navigation	✓	N.Q.
P16	Male	23	Answer	✓	✓
P17	Male	22	Chatbot	✓	✓
P18	Female	24	Balanced	N.Q.	✓
P19	Female	22	Chatbot	✓	✓
P20	Male	22	Navigation	✓	✓
P21	Male	24	Answer	N.Q.	✓
P22	Female	28	Navigation	✓	✓
P23	Female	22	Navigation	N.Q.	N.Q.
P24	Male	20	Chatbot	N.Q.	N.Q.

the Baseline system design, and the survey instruments. We conducted a within-subject study to eliminate the potential influence of the participants' personal particulars, *e.g.,* argumentative writing capabilities [124].

5.1 Study Procedures

With the approval of our institution's IRB, we recruited 24 CQA lurkers (ten female, fourteen male; age range 20-29; CQA usage frequency seven daily, twelve 4-6 days a week, five at least once a week; all composing no more than five posts over the past year on CQA platforms; summarized in Table 4) via online advertising, social media, and word-of-mouth at local universities. The inclusion criteria were that participants are interested in the given topics and have moderate knowledge of them. The participants were asked to interact with two sets of collective arguments with roughly equal information load (Table 2), fill out questionnaires about the interaction experience, and finally elaborate on their questionnaire responses in a short exit interview. The interaction with each system lasted around 10-40 minutes, and other parts of the study (*i.e.,* initial briefing and final interview) lasted around 30 minutes. Each participant received a $20 gift card for completing the study.

To reduce the Hawthorne effect (*a.k.a.,* observer or experimenter effect) [10, 73, 121], we inquired about participants' interest and

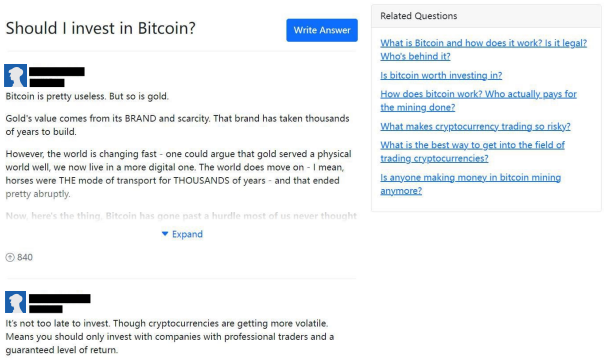

Figure 5: The Quora-like CQA Baseline System.

knowledge of other topics (*e.g.,* gun controls) together with those to experiment in the pre-screening questionnaire. After briefing the context, the task, and features of the Baseline system or CoArgue with a non-experiment topic, the participants were asked to interact with the designated system freely without the presence of the researchers. They received links to our systems at a random non-working time; they could start the experiment sessions anytime online and complete them without time limits. We also ensured that they did not receive the link to the second session until at least one day after finishing the first session. We selected two relatively popular experimental topics to avoid introducing knowledge prerequisite as the confounding factor. Four combinations were formed to counterbalance the experiment using *Latin Square*:

- Self-driving Car (Baseline) - Bitcoin Investment (CoArgue)
- Bitcoin Investment (Baseline) - Self-driving Car (CoArgue)
- Bitcoin Investment (CoArgue) - Self-driving Car (Baseline)
- Self-driving Car (CoArgue) - Bitcoin Investment (Baseline)

Upon completing the study, we examined the novel argumentative output generated by participants. Specifically, we extracted claims and premises (section 4.1.1) from the participants' writing and then compared them with those from the existing answer posts using the Jaccard index [80]. All but one claim from one participant (a paraphrase of an existing claim with a similarity score of .846) were identified as the novel. Among the novel argumentative elements identified from user output, the maximum similarity score was .364 for claims and .294 for premises. After the similarity check, we marked participants' writing output with no argumentative elements (*i.e.,* Not Qualified) for further analysis (Table 4).

5.2 The Quora-like CQA Baseline System

We built a Baseline system that replicated Quora in functionality but resembled CoArgue in UI components and styles so as to reduce the possible influence of other variables [44]. The Quora-like baseline system (Figure 5) excluded all the functionalities of the Answer View, leaving all answers ranking by upvotes. The Navigation and Chatbot Views were replaced by the original *related question* list of the corresponding Quora question-answer threads. Users could browse answers and the associated metadata (*e.g.,* upvotes, author profiles) or navigate to other related questions.

5.3 Hypotheses and Measurements

Following the evaluation pipeline of interactive systems [129, 134], we assessed the systems in three aspects: system usefulness, visual design & interactions, and system usability. We surveyed the related works to construct the questionnaire for the evaluation and correspondingly derive the hypotheses.

System usefulness. Referring to the evaluation methods of the behavior change supporting systems [84], we asked about the perceived support in motivation [19] and ability [29] respectively as part of the user experience (*H1a-b*). Moreover, as revealed in the Formative Study and previous works on computer-aided systems [44, 93], engagement is one of the key constructs of the user experience (*H1c*). We thus measured engagement based on Brien's theoretical model [82], which addressed the positive experience defined by the flow theory [17]. For the outcome analysis, we inquired the participants on their confidence [15] and self-perceived quality [124] of their writings (*H2a-b*). Apart from the subjective evaluation, we also objectively examined answer posts created by them from both systems (*H2c-g*). Specifically, we measured the readability [28], length, number of claims, number of premises, and their sum identified by our pipeline (section 4.1.1).

Visual Designs & Interactions. We firstly measured both systems' intuitiveness [133, 134], which means if the system is intuitive to use (*H3a-b*). Following practice of [134], in the final interview, we asked the participants to rate the usefulness of different Views of our system on the ability support, motivation support, and engagement to further understand how to support lurkers in contributing to collective arguments with visual designs & interactions.

System Usability. The trade-off between functionality and usability is a common issue for computer-supported systems [34, 44]. In our evaluation, we measured it with reference to [133, 134], a shorter version of the standard System Usability Scale (SUS) questionnaire [9] (*H4a-c*).

To conclude with, we derived the following hypotheses:

- **H1** Compared to the Baseline, CoArgue is more effective in supporting lurkers' motivation (*H1a*) and ability (*H1b*), plus improving their engagement (*H1c*) during the contribution process to the collective arguments.
- **H2** Compared to the Baseline, CoArgue can help lurkers to improve their confidence (*H2a*) on and self-perceived quality (*H2b*) of their contributions. The composed answer posts with CoArgue are also higher in readability (*H2c*), longer in length (*H2d*), and with more fruitful argumentative content (*H2e-g*).
- **H3** Compared to the Baseline, CoArgue's visual design (*H3a*) and the interaction (*H3b*) are more intuitive to users.
- **H4** Compared to the Baseline, CoArgue is perceived with improved performance on "easy to use" (*H4a*), "easy to learn" (*H4b*), and "willing to use again" (*H4c*).

6 RESULTS

We obtained the participants' ratings on the motivation & ability support and engagement level during the collective arguments creation process, confidence and self-perceived quality towards the output writings, and the intuitiveness & usability of the interaction design with CoArgue and the Baseline system, respectively. All tests were measured with a 7-point Likert scale, with 1 being the most negative impression (*e.g.,* not useful at all) and 7 being the most positive impression (*e.g.,* very useful).

We performed Wilcoxon signed-rank test [132] to assess the difference in the participants' ratings and paired sample t-test [99] to compare their writing outcomes concerning various factors of the two systems. For paired sample t-test, we verified the normality assumption with Shapiro-Wilk test [107] where each test variable received a p-value greater than 0.05. Hence, the normality assumption holds. We also confirmed that there was no significant effect regarding the topic assignment and the presentation order of the two study conditions (Wilcoxon signed-rank test, with all $p > 0.05$). Under the current sample size, all hypotheses of **H1** were accepted with a power level [26] greater than 0.9, while other accepted hypotheses achieved a power level of at least 0.75. Hence, the sample size was adequate for conducting the statistical tests.

In the rest of this section, we evaluated our system from usefulness, the efficacy of visual and interaction designs, and the overall usability. We supported our findings with statistical inference and supplemented it with qualitative reflection from the participants. Table 4 lists the demographics of the participants. Following [103, 117, 123], we calculated the test statistic, p-value, and corresponding effect size[10] for each hypothesis (Table 5).

6.1 System Usefulness

We evaluate the effectiveness of CoArgue on supporting collective argument composition from two aspects: subjective user experience ratings and task outcomes.

6.1.1 User Experience. During the interaction process, participants felt significantly more supported in terms of their motivation ($W = 223.50, p < 0.001$) and ability ($W = 206.5, p < 0.001$) when using CoArgue , compared to the Baseline system; *H1a* and *H1b* were accepted. Moreover, they were also significantly more engaged in the contributing process with CoArgue than with the Quora-like baseline system ($W = 260.50, p < 0.001$); *H1c* was accepted, which means **H1** was fully accepted.

Overall, over half of the participants (14/24) explicitly expressed that CoArgue could help them create answer posts efficiently, as they had a high-level overview of the argumentative elements, which made various opinions presented in the collective arguments easy to follow. Details of how CoArgue 's interaction design supported **H1** are analyzed in Section 6.2.2. In addition, we observed a large variance in participants' interaction time with the systems (10-40 minutes). Some participants only read a few high-voted answer posts and felt that these discussions were comprehensive enough with little room for anything new, while other participants spent extensive efforts to compose detailed answer posts.

[10] In hypothesis testing, effect size is the objective and standardized measure of the size of a particular effect [123]. Different hypothesis testing methods may choose different coefficients as effect size and have different thresholds for interpreting effect size. T-test conventional effect sizes, proposed by Cohen, are Cohens' d with thresholds: 0.2 -< 0.5 (small effect), 0.5 -< 0.8 (moderate effect) and >= 0.8 (large effect). For Wilcox-signed rank test, recommended values are (Rank-Biserial Correlation): 0.10 - < 0.3 (small effect), 0.30 - < 0.5 (moderate effect) and >= 0.5 (large effect) [45].

Table 5: The statistical analysis of user feedback with Baseline and CoArgue, where the p-value (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) is reported. H2c-g were analyzed using paired sample t-test, while others using Wilcoxon signed rank test. Effect size with large or moderate magnitude is *highlighted*.

Category	Factor	Baseline Mean(S.D.)	CoArgue Mean(S.D.)	Statistics					Hypothesis
				W	T	p-value	Sig.	Eff. Size	
Experience	Motivation	2.75(1.51)	5.21(1.18)	223.50		<0.001	***	***0.80***	H1a acc.
	Ability	3.38(1.47)	5.42(1.41)	206.50		<0.001	***	***0.81***	H1b acc.
	Engagement	25.54(9.52)	37.25(5.55)	260.50		<0.001	***	***0.77***	H1c acc.
Outcome (Subjective)	Confident	3.88(1.65)	5.12(0.90)	163.50		0.003	**	***0.57***	H2a acc.
	Self-perceived quality	3.38(1.53)	4.54(1.28)	128.50		0.007	**	***0.50***	H2b acc.
Outcome (Objective)	Readability	59.63(10.01)	63.08(9.37)		1.45	0.084	+	0.36	H2c rej.
	# Words	136.56(77.39)	158.81(76.13)		1.03	0.159	-	0.26	H2d rej.
	# Claims	2.31(1.08)	2.69(0.95)		0.97	0.173	-	0.24	H2e rej.
	# Premises	3.62(2.06)	4.88(2.00)		2.13	0.025	*	***0.53***	H2f acc.
	# Claims + # Premises	5.94(2.93)	7.56(2.16)		2.06	0.029	*	***0.51***	H2g acc.
Intuitiveness	Visual design	5.33(1.58)	6.08(0.83)	125.00		0.043	*	***0.30***	H3a acc.
	Interactions	4.96(1.71)	5.71(1.30)	121.00		0.062	+	0.28	H3b rej.
Usability	Easy to use	6.42(0.83)	5.50(1.47)	15.00		0.009	**	***0.47***	H4a rej.
	Easy to learn	6.17(1.09)	5.54(1.44)	46.00		0.075	+	***0.33***	H4b rej.
	Willing to use again	4.25(1.70)	5.54(1.35)	169.00		0.008	**	***0.52***	H4c acc.

6.1.2 Outcome Analysis. As shown in Table 4, nine participants did not compose qualified answer posts with the Baseline system, while seven participants did not do so with CoArgue. Although the contribution rate in the CoArgue condition (17/24) was only slightly higher than that of the Baseline condition (15/24), the quality and the pattern of participants' contributions could vary. Therefore, in the following part, we further inspected the answer posts written by participants who actually joined the collective arguments.

Subjective Evaluation. Compared to the Baseline system, the participants were significantly more confident about the answer post composed with CoArgue ($W = 163.50, p = 0.003$), and perceived it to be of a significantly higher quality ($W = 128.50, p = 0.007$); thus, *H2a* and *H2b* were accepted.

Objective Evaluation. We applied paired sample t-test to evaluate the writings objectively, as described in section 5.3. Results indicated that there was no significant difference in the readability (*H2c*) and the word count (*H2d*) of the participants' contributions between the two conditions; therefore, both hypotheses were rejected. Yet, the participants wrote significantly more premises ($T = 2.13, p = 0.025$) and produced significantly more claims and premises together ($T = 2.06, p = 0.029$) with CoArgue; but there was no significant difference in terms of claims alone. This means *H2f* and *H2g* were accepted, while *H2e* was rejected. As such, **H2** was partially accepted.

Contribution Patterns. Apart from the difference in the number of claims and premises produced by the participants, as shown in Table 5, we further analyzed the contribution patterns of the participants by processing the resulting collective arguments (existing answer posts plus those created by participants) of the two systems and the two topics (*i.e.,* bitcoin investment and self-driving car), respectively, using the claim centers clustering algorithm described in section 4.1.2. As the clustered claim centers did not have the one-to-one mapping relationship due to the difference in the participants' answer posts as part of the input, we analyzed the statistics holistically on the system level (illustrated in Table 6). We also calculated the ratio of new claims written by participants for each claim center. The average ratio of CoArgue was 2.9% higher with a larger S.D. than that of the Baseline system. These data might indicate that participants using CoArgue were likely to explore and contribute more to claim centers less visited by the previous contributors.

Table 6: Statistics of the clustered claim centers with participants' contribution (*C.C.* = Claim Center).

	CoArgue	Baseline
# Claim Centers of Bitcoin Investment	16	16
# Claim Centers of Self-driving Car	13	12
Avg. ratio of new Claims in C.C.	21.1%	18.2%
S.D. on ratios of new Claims in C.C.	0.246	0.218

Despite the significant results, we could see that the magnitude (*i.e.,* the range of effect size of the associated tests) of improvements in the three aforementioned evaluation aspects has a decreasing trend: the user perceived support during interaction (*H1a-b*: *large*), subjective perception of the outcome (*H2a-b*: *moderate* to *large*), the corresponding objective evaluation (*H2c-g*: *small* to *moderate*). This finding implied that people might feel rather motivated and supported during the interaction process, but the effect would be weakened when they subjectively assess their own outcomes and further depleted in their actual performance, reflecting the impediments lurkers might face in changing their behavior. As pointed out by P24, who did not write anything in both systems, *"although my eager to write something is somehow stronger with CoArgue [compared with the Baseline], it still could not reach my threshold [to actually write something]."*

6.2 Efficacy of Visual and Interaction Designs

6.2.1 Intuitiveness. For participants, visual designs and interactions of both systems were intuitive, with mean ratings around 5-6/7 (Table 5). Compared to the Baseline, the visual designs of CoArgue was considered to be significantly more intuitive ($W = 125.00, p = 0.043$) but the effect was relatively small (*H3a* accepted); the intuitiveness of its interaction was only marginally higher ($W = 121.00, p = 0.062$); *H3b* marginally accepted. Paralleled to the quantitative results, five participants (P2-3, P5, P12, P21) explicitly praised that the user interface of CoArgue was simple and neat in design, given its functionalities.

6.2.2 User Perceptions on Design. To further analyze which features of CoArgue contributed to the change of lurking behavior, we invited participants to give their ratings (7-point Likert scale) on individual Views of CoArgue regarding their support on motivation, ability, and engagement. Figure 6 presents the results in details.

In terms of the effect on boosting motivation, the Navigation View gained the most significant number of *strongly agree* and *agree* (ten in total). At the same time, the Contribution View received the most number of *strongly agree* (five compared to one or two in other Views). In the post-study interview, two participants said they were more willing to start a new post because the CoArgue organized information well (Navigation View) for them (P5, P18). Another four participants added that the Contribution View encouraged them to refine their initial drafts of answer posts by showing detailed statistical data about their contributions (P1-2, P7-8).

Regarding the ability support, the Navigation View received predominate approval from the participants with 21/24 *strongly agree* and *agree*, followed by the Answer View (15/24 and no negative ratings). In particular, seven participants considered the Navigation View and Answer View formed a great synergy to help them get started in writing (P1-2, P5-6, P11, P16, P22). Specifically, they explored different stances in the Navigation View first, then dived into the claim centers they were interested in, and later jumped to the original discussion threads with the corresponding argumentative elements highlighted in the Answer View.

For the source of the engagement, participants who benefit from the synergy between the Answer View and Navigation View (as mentioned in the last paragraph) were likely to consider the two Views helpful to their engagement with CoArgue. As for the Contribution View, three participants acknowledged that they gained self-fulfillment with this View, which eventually added to their engagement during the interaction with CoArgue (P6, P11, P17). By comparison, participants seemed to hold diverse opinions about the Chatbot View. We found that participants who usually enjoy conversing with chatbots engaged more during the interaction with the Chatbot View (P12, P19, P24). In contrast, those who do not like to have free conversations with chatbots would regard this feature as simply repeating some pre-defined scripts and feel bored about it (P5, P13, P23).

6.3 System Usability

Generally speaking, participants perceived CoArgue as significantly more complicated than the Baseline system (*H4a* and *H4b* rejected with reversed significance, see Table 5), but they were significantly more willing to use it again ($W = 169.00, p = 0.008$; *H4c* accepted). Regarding specific dimensions of usability, CoArgue achieved a rating of $Mean = 5.50(SD = 1.47)$ in *easy to use* and $Mean = 5.54(SD = 1.44)$ in *easy to learn*, indicating good usability with an average rating over 5/7 in the 7-point Likert scale, yet lower than the Baseline that rated as $Mean = 6.42(SD = 0.83)$ and $Mean = 6.17(SD = 1.09)$ respectively.

As expected, CoArgue with more additional features are considered significantly less easy to use and learn compared to the Baseline system. Four participants explicitly mentioned the concern on its learning cost (P2, P4, P16, P24), and P16 added that *"the system is full-functioning, but some of the features are redundant [for me], and I do not wish to learn [all of the features]"*. Nevertheless, although the reduced usability compared to the Baseline, most participants (18/24) still expressed their willingness to use CoArgue again in the post-study interview, which means the decline in usability was largely acceptable.

7 DISCUSSION

Lurking is a common practice in a wide range of online communities [13, 46, 51, 58, 62], and the underlying reasons are often complex [46, 58, 62]. Our Formative Study revealed that lurkers in CQA platforms face various obstacles, ranging from confidence and willingness to the ability to digest and write arguments. Previous works on either supporting CQA digestion [64], argument reading [49, 50], or argument writing [125, 134] only addressed one facet of these challenges. Our work demonstrates the effectiveness of a comprehensive de-lurking approach that supports CQA content digestion, argument processing, and confidence building in collective argument writing.

The rest of this section presents the design considerations (**DC**s) we derived from the study results for future construction of tools like CoArgue. We further summarized the limitations and the future works.

7.1 Design Considerations

As an overview, **DC1** is related to the design of motivational support, while **DC2 & DC3** concern ability, and finally, **DC4** is a reflection on engagement.

*7.1.1 Provide More Timely Feedback on Users' Contribution (**DC1**).* Despite the simplicity of the Contribution View design, by seeing the statistics and the highlights of their writing, the participants still regarded it to help motivate them to make contributions to collective arguments (out of 24 participants, five rated *strongly agree*, in contrast to at most two for other Views, Figure 6). The participants considered that such feedback on their posts could *"let me know how good my answer post was among the existing ones"* (P5), supported self-fulfillment (P6, P22), and gave incentives to achieve *"a better-looking statistic"* (P11, 18). In fact, in the Formative Study, four participants mentioned that not receiving feedback after submitting the answer post is one of the critical problems of the mainstream CQA platforms, as it would make users feel that no one values their contributions. Existing works also have demonstrated that the timely response to messages sent by the members is critical to the sustainable development of the online communities [24, 53, 54, 72]. Therefore, considering the loosely connected nature of the CQA communities [43, 113], prompt feedback by the system

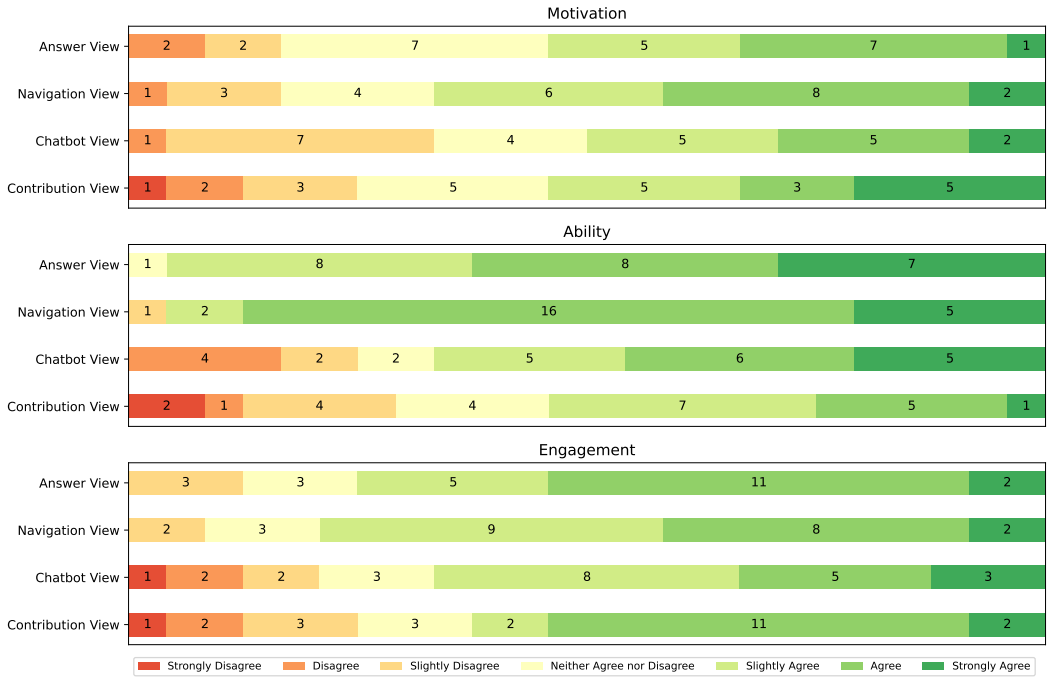

Figure 6: Participants' ratings regarding the usefulness of the four Views towards motivation, ability, and engagement.

regarding the connection of the answer post created to existing posts, similar to the Contribution View of CoArgue, is a good option for mainstream CQA platforms to keep users motivated. Yet it is beneficial also to take the accuracy of algorithms to be applied into consideration (**DC3**). Moreover, P17 suggested that the feedback of CoArgue should focus on the writing itself, *i.e.,* highlights of the argumentative elements, as *"it may bring peer pressure to users with 'Player versus Player' type of feedback or discourage them with the unsatisfactory results, especially for lurkers"*. Therefore, the type of feedback is also worth thinking about.

*7.1.2 Provide a Holistic Summary of Existing Collective Arguments (**DC2**).* From the Formative Study, we found it challenging for users to digest an overwhelming number of answer posts to distill the collective arguments. Providing a holistic overview of the existing contents is thus beneficial for grasping argumentative elements from many unstructured online discussions. Prior studies' findings inspire our Navigation View design that generating clear outlines is an effective way to facilitate people to process and understand large-scale UGC [14, 57, 104, 122]. User study Results suggested that CoArgue demonstrated a more robust capability of retaining user interest in the question thread they were reading compared to the Baseline condition. Moreover, nearly half of the participants (11/24) also acknowledged the effectiveness of the summary powered by Navigation View in supporting critical thinking and breaking the echo chamber. For instance, P22 considered the interaction with the Navigation View useful for continuously updating her knowledge of the current context, and P19 said that such a process helped her generate some novel insights. In brief, a holistic summary of the existing collective arguments allows users of CQA to develop a good understanding of the given topic and the current arguments established around it, which lays an essential foundation for further participation in the discussions if any.

*7.1.3 Calibrate Algorithm Accuracy in Identifying Argumentative Elements (**DC3**).* Existing research indicated that the model performance of the AI systems is crucial to the success of human-AI collaborations [21, 110, 135]. In the post-study interview, around one-third of participants (9/24) explicitly mentioned their concern about the accuracy of the AI algorithms, which compromised their trust to a different extent during the interaction with CoArgue. According to the participants, it was relatively acceptable to see both positive and negative claim centers being classified to the neutral category (P14), or simply receive inconsistent dialogue flow from the chatbot (P9). By contrast, mislabeling positive claim centers as negative or vice versa was more intolerable. Users who encountered such cases might have a high chance of giving up on the Navigation View (P3, P11, P13). The most negative impression came from system feedback that incorrectly predicted the stance of or extracted wrong claims and premises from the users' writing (P23). Users got extraordinarily frustrated and quit participating, feeling that the system would not recognize their contributions appropriately. In summary, user expectations towards AI accuracy of different features and in different usage scenarios vary greatly, which is also found in previous research [77, 86, 90]. Future systems designers like CoArgue must carefully calibrate the expected functionalities and the algorithm's accuracy.

*7.1.4 Maintain User Engagement with Alternative Interactive Technologies (**DC4**).* Maintaining user engagement during the process is essential for computer-aided systems that support users in content

creation. [22, 44, 69]. In CoArgue, the Chatbot View provided an alternative way to interact with the collective arguments, especially when participants get bored with the Navigation View. Aligned with previous works on chatbot interactions [65, 70, 78, 108], participants generally confirmed that the chatbot's proactive behaviors helped maintain their involvement in the process. P3 and P19 started chatting with the chatbot as they felt like being invited to the discussion, while P2 was simply curious about the chatbot and then found it fun to interact with. P4, P15, and P18, on the other hand, were impressed by the diligence of the chatbot in attempting to guide them to compose answer posts step by step. Despite the specific reasons, the incorporation of the chatbot function indeed augmented and enriched the interactive experiences in CQA, enhancing user engagement in both reading and writing collective arguments. Therefore, to design tools like CoArgue, different forms of interactions can be seamlessly integrated to balance the various individual preferences on control and interaction efficiency, aiming at improving user engagement without impairing usability.

7.2 Limitations and Future Work

Our design and studies have several limitations. First, the Chatbot View of CoArgue aims to facilitate the composition of answer posts, and the Navigation View is designed to give users an overall coverage of the existing stances and claims of the collective argument. Yet, as indicated by [64], users might abandon their original thoughts after inspecting this overview, thus being "over-guided" by the system. Second, our Formative Study and User Study participants were primarily young adults (age range 20-29), and this skewed age distribution could introduce some bias. More experiments with diverse user groups should be conducted to improve the accessibility and inclusiveness of CoArgue. Third, to avoid potential prejudice introduced by sensitive and controversial topics, *e.g.,* the overturn of *Roe v. Wade*, we selected technology-related topics (*i.e.,* bitcoin investment and self-driving cars) from those mentioned by Formative Study participants as the final evaluation context of CoArgue. Apart from these topics, we would also like to explore the efficacy of CoArgue on less popular topics (*e.g.,* those with relatively high knowledge prerequisites). In fact, we foresaw CoArgue's generalizability to a wide range of CQA topics, as the collective arguments processing pipeline we proposed in this work (section 4.1) is context-independent and free of human labeling. Fourth, besides the objective assessment of participants' writing output, we applied mostly subjective measurements to evaluate our system. As CoArgue should appear novel to first-time users, it might influence participants' feedback by novelty effect. Long-term field studies can be conducted in future works to mitigate this potential bias.

Finally, in the scope of this paper, we did not monitor the long-term effects of CoArgue on reducing CQA users' lurking behaviors. Therefore, although CoArgue demonstrated its capabilities of triggering lurkers to make more contributions in the short term, it is still unclear if the persistent application of CoArgue could reverse the lurking habit of users. A few participants commented on this issue in the post-study interview. P3 considered CoArgue *"a first aid kit designed for the platform,"* and thus she regarded the motivation support from the tool as only having a temporary impact on her. P1 and P22 postulated that reversing their lurking habits requires more concrete stimulants and mechanisms. Therefore, future research could evaluate the endurance of CoArgue's effects in countering lurking behaviors to transform CQA lurkers into frequent contributors to collective arguments permanently. Moreover, CoArgue can be further enhanced to expand the existing scope of interest of CQA users, lurkers included, in the long run.

8 CONCLUSION

In this work, we presented CoArgue, a proof-of-concept supporting tool to foster lurkers' contribution to collective arguments of CQA platforms. We designed an NLP pipeline to process the collective arguments and an interactive interface on top based on the Formative Study results. Compared to a Quora-like Baseline interface, a within-subject user study demonstrated that CoArgue significantly improved CQA lurkers' motivation, ability, and engagement in making contributions. CoArgue also significantly improved the quality of the answer posts composed by lurkers both subjectively and objectively. We further summarized design considerations to guide future work to design lurker supporting tools like CoArgue.

ACKNOWLEDGMENTS

Many thanks to the anonymous reviewers for their insightful suggestions. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under General Research Fund (GRF) with Grant No. 16204420.

REFERENCES

[1] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 58, 13 pages. https://doi.org/10.1145/3411764.3445683

[2] Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (2021), 102597.

[3] Steven Baker, Jenny Waycott, Romina Carrasco, Ryan M. Kelly, Anthony John Jones, Jack Lilley, Briony Dow, Frances Batchelor, Thuong Hoang, and Frank Vetere. 2021. Avatar-Mediated Communication in Social VR: An In-Depth Exploration of Older Adult Interaction in an Emerging Communication Platform. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 491, 13 pages. https://doi.org/10.1145/3411764.3445752

[4] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 258–266. https://aclanthology.org/2022.lrec-1.27

[5] Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, and Prasenjit Mitra. 2012. Thread specific features are helpful for identifying subjectivity orientation of online forum threads. In *Proceedings of COLING 2012*. 295–310.

[6] Alexander Bondarenko, Matthias Hagen, Martin Potthast, Henning Wachsmuth, Meriem Beloucif, Chris Biemann, Alexander Panchenko, and Benno Stein. 2020. Touché: First Shared Task on Argument Retrieval. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II* (Lisbon, Portugal). Springer-Verlag, Berlin, Heidelberg, 517–523. https://doi.org/10.1007/978-3-030-45442-5_67

[7] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly? Analyzing clarification questions in CQA. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 345–348.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[9] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability evaluation in industry* 189 (1996), 4–7.

[10] Felix Carros, Johanna Meurer, Diana Löffler, David Unbehaun, Sarah Matthies, Inga Koch, Rainer Wieching, Dave Randall, Marc Hassenzahl, and Volker Wulf. 2020. Exploring Human-Robot Interaction with the Elderly: Results from a Ten-Week Case Study in a Care Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376402

[11] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 685, 14 pages. https://doi.org/10.1145/3411764.3445068

[12] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 545 (nov 2022), 37 pages. https://doi.org/10.1145/3555603

[13] Fei-Ching Chen and Hsiu-Mei Chang. 2011. Do Lurking Learners Contribute Less? A Knowledge Co-Construction Perspective. In *Proceedings of the 5th International Conference on Communities and Technologies* (Brisbane, Australia) *(C&T '11)*. Association for Computing Machinery, New York, NY, USA, 169–178. https://doi.org/10.1145/2103354.2103377

[14] Rocio Chongtay, Mark Last, and Bettina Berendt. 2018. Responsive News Summarization for Ubiquitous Consumption on Multiple Mobile Devices. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 433–437. https://doi.org/10.1145/3172944.3172992

[15] Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (New York, NY, USA, 2012-02-11) *(CSCW '12)*. Association for Computing Machinery, 383–392. https://doi.org/10.1145/2145204.2145265

[16] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[17] Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row, New York, USA.

[18] Kushal Dave, Martin Wattenberg, and Michael Muller. 2004. Flash Forums and ForumReader: Navigating a New Kind of Large-Scale Online Discussion. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 232–241. https://doi.org/10.1145/1031607.1031644

[19] Roelof A.J. de Vries, Khiet P. Truong, Sigrid Kwint, Constance H.C. Drossaert, and Vanessa Evers. 2016. Crowd-Designed Motivation: Motivational Messages for Exercise Adherence Based on Behavior Change Theory. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016-05-07) *(CHI '16)*. Association for Computing Machinery, 297–308. https://doi.org/10.1145/2858036.2858229

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[21] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[22] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* 51, 5, Article 99 (nov 2018), 39 pages. https://doi.org/10.1145/3234149

[23] Lorik Dumani, Patrick J Neumann, and Ralf Schenkel. 2020. A framework for argument retrieval. In *European Conference on Information Retrieval*. Springer, 431–445.

[24] Rosta Farzan, Robert Kraut, Aditya Pal, and Joseph Konstan. 2012. Socializing Volunteers in an Online Community: A Field Experiment. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 325–334. https://doi.org/10.1145/2145204.2145256

[25] Haakon Faste and Honray Lin. 2012. *The Untapped Promise of Digital Mind Maps*. Association for Computing Machinery, New York, NY, USA, 1017–1026. https://doi.org/10.1145/2207676.2208548

[26] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[27] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An Intelligent Discussion-Bot for Answering Student Queries in Threaded Discussions. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (Sydney, Australia) *(IUI '06)*. Association for Computing Machinery, New York, NY, USA, 171–177. https://doi.org/10.1145/1111449.1111488

[28] Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education* (1943).

[29] Rachel L. Franz, Jacob O. Wobbrock, Yi Cheng, and Leah Findlater. 2019. Perception and Adoption of Mobile Accessibility Features by Older Adults Experiencing Ability Changes. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2019-10-24) *(ASSETS '19)*. Association for Computing Machinery, 267–278. https://doi.org/10.1145/3308561.3353780

[30] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[31] Shogo Fujita, Tomohide Shibata, and Manabu Okumura. 2020. Diverse and Non-redundant Answer Set Extraction on Community QA based on DPPs. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5309–5320.

[32] Andrew Gambino and S. Shyam Sundar. 2019. Acceptance of Self-Driving Cars: Does Their Posthuman Ability Make Them More Eerie or More Desirable?. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312870

[33] S Geerthik, S Venkatraman, and K Rajiv Gandhi. 2016. Reward rank: A novel approach for positioning user answers in community question answering system. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)*. IEEE, 1–6.

[34] Nancy C. Goodwin. 1987. Functionality and Usability. *Commun. ACM* 30, 3 (March 1987), 229–233. https://doi.org/10.1145/214748.214758

[35] Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*. Springer, 287–299.

[36] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[37] Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2010. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences* 180, 19 (2010), 3686–3702.

[38] F Maxwell Harper, Daniel Moy, and Joseph A Konstan. 2009. Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the sigchi conference on human factors in computing systems*. 759–768.

[39] Joe E Heimlich and Nicole M Ardoin. 2008. Understanding behavior to understand behavior change: A literature review. *Environmental education research* 14, 3 (2008), 215–237.

[40] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. 11–21.

[41] Enamul Hoque and Giuseppe Carenini. 2015. ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 169–180. https://doi.org/10.1145/2678025.2701370

[42] Enamul Hoque, Shafiq Joty, Luis Marquez, and Giuseppe Carenini. 2017. CQAVis: Visual Text Analytics for Community Question Answering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3025171.3025210

[43] Shafiq Joty, Alberto Barrón-Cedeno, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2019. Global thread-level inference for comment classification in community question answering. *arXiv preprint arXiv:1911.08755* (2019).

[44] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting Visual Metaphor Ideation through Multi-Dimensional Example-Based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 427, 15 pages. https://doi.org/10.1145/3411764.3445325

[45] Alboukadel Kassambara. 2021. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://CRAN.R-project.org/package=rstatix R package version 0.7.0.

[46] Avleen Kaur, C Estelle Smith, and Loren Terveen. 2021. Sway Together, Stay Together: Visualizing Spiritual Support Networks Through the SoulGarden Prototype. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 84–88.

[47] Imrul Kayes, Nicolas Kourtellis, Francesco Bonchi, and Adriana Iamnitchi. 2015. Privacy Concerns vs. User Behavior in Community Question Answering. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (Paris, France) *(ASONAM '15)*. Association for Computing Machinery, New York, NY, USA, 681–688. https://doi.org/10.1145/2808797.2809422

[48] Irni Eliana Khairuddin and Corina Sas. 2019. An Exploration of Bitcoin Mining Practices: Miners' Trust Challenges and Motivations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland

Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300859

[49] Dana Khartabil, Christopher Collins, S Wells, Benjamin Bach, and Jessie Kennedy. 2021. Design and Evaluation of Visualization Techniques to Facilitate Argument Exploration. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 447–465.

[50] Dana Khartabil, S Wells, and Jessie Kennedy. 2016. Large-scale Argument Visualization (LSAV).. In *EuroVis (Posters)*. 65–67.

[51] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 87 (apr 2021), 26 pages. https://doi.org/10.1145/3449161

[52] Maximilian Klein, Jinhao Zhao, Jiajun Ni, Isaac Johnson, Benjamin Mako Hill, and Haiyi Zhu. 2017. Quality Standards, Service Orientation, and Power in Airbnb and Couchsurfing. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 58 (dec 2017), 21 pages. https://doi.org/10.1145/3134693

[53] Yubo Kou and Colin M. Gray. 2017. Supporting Distributed Critique through Interpretation and Sense-Making in an Online Creative Community. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 60 (dec 2017), 18 pages. https://doi.org/10.1145/3134695

[54] Yubo Kou and Colin M. Gray. 2018. Towards Professionalization in an Online Community of Emerging Occupation: Discourses among UX Practitioners. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork* (Sanibel Island, Florida, USA) *(GROUP '18)*. Association for Computing Machinery, New York, NY, USA, 322–334. https://doi.org/10.1145/3148330.3148352

[55] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[56] Hui-Min Lai and Tsung Teng Chen. 2014. Knowledge sharing in interest online communities: A comparison of posters and lurkers. *Computers in Human Behavior* 35 (2014), 295–306.

[57] Heidi Lam and Patrick Baudisch. 2005. Summary Thumbnails: Readable Overviews for Small Screen Web Browsers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 681–690. https://doi.org/10.1145/1054972.1055066

[58] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to Participate in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1927–1936. https://doi.org/10.1145/1753326.1753616

[59] Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. *arXiv preprint arXiv:2006.00843* (2020).

[60] Long T Le, Chirag Shah, and Erik Choi. 2017. Bad users or bad content? Breaking the vicious cycle by finding struggling students in community question-answering. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 165–174.

[61] 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*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[62] Ang Li, Zheng Yao, Diyi Yang, Chinmay Kulkarni, Rosta Farzan, and Robert E Kraut. 2020. Successful online socialization: lessons from the Wikipedia education program. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–24.

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

[64] Chengzhong Liu, Zeyu Huang, Dingdong Liu, Shixu Zhou, Zhenhui Peng, and Xiaojuan Ma. 2022. PlanHelper: Supporting Activity Plan Construction with Answer Posts in Community-Based QA Platforms. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 454 (nov 2022), 26 pages. https://doi.org/10.1145/3555555

[65] Chengzhong Liu, Shixu Zhou, Yuanhao Zhang, Dingdong Liu, Zhenhui Peng, and Xiaojuan Ma. 2022. Exploring the Effects of Self-Mockery to Improve Task-Oriented Chatbot's Social Intelligence. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1315–1329. https://doi.org/10.1145/3532106.3533461

[66] Qiaoling Liu, Tomasz Jurczyk, Jinho Choi, and Eugene Agichtein. 2015. Real-Time Community Question Answering: Exploring Content Recommendation and User Notification Strategies. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 50–61. https://doi.org/10.1145/2678025.2701392

[67] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 272, 14 pages. https://doi.org/10.1145/3411764.3445233

[68] Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and summarizing answers in community-based question answering services. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*. 497–504.

[69] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 313, 25 pages. https://doi.org/10.1145/3491102.3517482

[70] Xiaojuan Ma, Emily Yang, and Pascale Fung. 2019. Exploring Perceived Emotional Intelligence of Personality-Driven Virtual Agents in Handling User Challenges. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1222–1233. https://doi.org/10.1145/3308558.3313400

[71] Jakub Macina, Ivan Srba, Joseph Jay Williams, and Maria Bielikova. 2017. Educational question routing in online student communities. In *Proceedings of the eleventh ACM conference on recommender systems*. 47–55.

[72] Diane Maloney-Krichmar and Jenny Preece. 2005. A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 201–232. https://doi.org/10.1145/1067860.1067864

[73] Rob McCarney, James Warner, Steve Iliffe, Robbert Van Haselen, Mark Griffin, and Peter Fisher. 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC medical research methodology* 7, 1 (2007), 1–8.

[74] Leland McInnes. 2018. Performance comparison of dimension reduction implementations. https://umap-learn.readthedocs.io/en/latest/performance.html

[75] Leland McInnes. 2018. Using UMAP for clustering. https://umap-learn.readthedocs.io/en/latest/clustering.html

[76] David W McMillan. 1996. Sense of community. *Journal of community psychology* 24, 4 (1996), 315–325.

[77] Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 146 (dec 2020), 22 pages. https://doi.org/10.1145/3432193

[78] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal, and Rosalind Picard. 2020. Promoting Wellbeing with Sunny, a Chatbot That Facilitates Positive Messages within Social Groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383062

[79] Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. *arXiv preprint arXiv:2011.01589* (2020).

[80] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, Vol. 1. 380–384.

[81] Blair Nonnecke and Jenny Preece. 2000. Lurker Demographics: Counting the Silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 73–80. https://doi.org/10.1145/332040.332409

[82] Heather O'Brien. 2016. *Theoretical Perspectives on User Engagement*. Springer International Publishing, Cham, 1–26. https://doi.org/10.1007/978-3-319-27446-1_1

[83] Heather L. O'Brien and Elaine G. Toms. 2008. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *J. Am. Soc. Inf. Sci. Technol.* 59, 6 (apr 2008), 938–955.

[84] Harri Oinas-Kukkonen. 2013. A foundation for the study of behavior change support systems. *Personal and ubiquitous computing* 17, 6 (2013), 1223–1235.

[85] Nigini Oliveira, Michael Muller, Nazareno Andrade, and Katharina Reinecke. 2018. The Exchange in StackExchange: Divergences between Stack Overflow and Its Culturally Diverse Participants. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 130 (nov 2018), 22 pages. https://doi.org/10.1145/3274399

[86] Thomas Olsson. 2014. Layers of User Expectations of Future Technologies: An Early Framework. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 1957–1962. https://doi.org/10.1145/2559206.2581225

[87] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 2016. Novelty based ranking of human answers for community questions. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in*

Information Retrieval. 215–224.

[88] Ina O'Murchu, John G Breslin, and Stefan Decker. 2004. Online Social and Business Networking Communities.. In *ECAI Workshop on Application of Semantic Web Technologies to Web Communities*, Vol. 107.

[89] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. *Supporting Comment Moderators in Identifying High Quality Online News Comments*. Association for Computing Machinery, New York, NY, USA, 1114–1125. https://doi.org/10.1145/2858036.2858389

[90] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-Shared AI Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376450

[91] Sharoda A Paul, Lichan Hong, and Ed H Chi. 2012. Who is authoritative? understanding reputation mechanisms in quora. *arXiv preprint arXiv:1204.3724* (2012).

[92] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[93] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring interactive question prompts for critical paper reading. 167 (2022), 102898. https://doi.org/10.1016/j.ijhcs.2022.102898

[94] Laura R. Pina, Carmen Gonzalez, Carolina Nieto, Wendy Roldan, Edgar Onofre, and Jason C. Yip. 2018. How Latino Children in the U.S. Engage in Collaborative Online Information Problem Solving with Their Families. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 140 (nov 2018), 26 pages. https://doi.org/10.1145/3274409

[95] Pavlin G Poličar, Martin Stražar, and Blaž Zupan. 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv* (2019), 731877.

[96] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[97] S. Rafaeli, G. Ravid, and V. Soroka. 2004. De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the.* 10 pp.–. https://doi.org/10.1109/HICSS.2004.1265478

[98] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[99] Amanda Ross and Victor L Willson. 2017. Paired samples T-test. In *Basic and advanced statistical tests*. Springer, 17–19.

[100] Dana Rotman, Jennifer Golbeck, and Jennifer Preece. 2009. The Community is Where the Rapport is – on Sense and Structure in the Youtube Community. In *Proceedings of the Fourth International Conference on Communities and Technologies* (University Park, PA, USA) *(C&T '09)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/1556460.1556467

[101] Pradeep K Roy, Jyoti P Singh, Abdullah M Baabdullah, Hatice Kizgin, and Nripendra P Rana. 2018. Identifying reputation collectors in community question answering (CQA) sites: Exploring the dark side of social media. *International Journal of Information Management* 42 (2018), 25–35.

[102] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300587

[103] Tetsuya Sakai. 2017. The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25–34.

[104] Amit Sarkar and G. Srinivasaraghavan. 2018. Contextual Web Summarization: A Supervised Ranking Approach. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 105–106. https://doi.org/10.1145/3184558.3186951

[105] Christian Severin Sauer and Thomas Roth-Berghofer. 2012. Solution Mining for Specific Contextualised Problems: Towards an Approach for Experience Mining. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/2187980.2188193

[106] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in Qualitative Research: Exploring Its Conceptualization and Operationalization. 52, 4 (2018), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8 pmid:29937585

[107] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.

[108] Donghoon Shin, Subeen Park, Esther Hehsun Kim, Soomin Kim, Jinwook Seo, and Hwajung Hong. 2022. Exploring the Effects of AI-Assisted Emotional Support Processes in Online Mental Health Community. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 300, 7 pages. https://doi.org/10.1145/3491101.3519854

[109] Yoshiyuki Shoji, Sumio Fujita, Akira Tajima, and Katsumi Tanaka. 2015. Who stays longer in community qa media?-user behavior analysis in cqa. In *International Conference on Social Informatics*. Springer, 33–48.

[110] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[111] Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing answers in non-factoid community question-answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 405–414.

[112] Ivan Srba and Maria Bielikova. 2015. Askalot: community question answering as a means for knowledge sharing in an educational organization. In *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing*. 179–182.

[113] Ivan Srba and Maria Bielikova. 2016. A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web (TWEB)* 10, 3 (2016), 1–63.

[114] Ivan Srba, Marek Grznar, and Maria Bielikova. 2015. Utilizing non-qa data to improve questions routing for users with low qa activity in cqa. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*. 129–136.

[115] Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 46–56.

[116] Victor J Strecher, Brenda McEvoy DeVellis, Marshall H Becker, and Irwin M Rosenstock. 1986. The role of self-efficacy in achieving health behavior change. *Health education quarterly* 13, 1 (1986), 73–92.

[117] Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the P value is not enough. *Journal of graduate medical education* 4, 3 (2012), 279–282.

[118] Na Sun, Patrick Pei-Luen Rau, and Liang Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior* 38 (2014), 110–117.

[119] Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. 2020. Frameworks for collective intelligence: A systematic literature review. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–36.

[120] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. 2013. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22nd international conference on World Wide Web*. 1249–1260.

[121] Feng Tian, Xiangmin Fan, Junjun Fan, Yicheng Zhu, Jing Gao, Dakuo Wang, Xiaojun Bi, and Hongan Wang. 2019. What Can Gestures Tell? Detecting Motor Impairment in Early Parkinson's from Common Touch Gestural Interactions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300313

[122] Sunny Tian, Amy X. Zhang, and David Karger. 2021. A System for Interleaving Discussion and Summarization in Online Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 241 (jan 2021), 27 pages. https://doi.org/10.1145/3432940

[123] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.

[124] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[125] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: an adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[126] Baoxun Wang, Xiaolong Wang, Cheng-Jie Sun, Bingquan Liu, and Lin Sun. 2010. Modeling semantic relevance for question-answer pairs in web social communities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 1230–1238.

[127] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2013. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the*

22nd international conference on World Wide Web. 1341–1352.

[128] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. TIARA: A Visual Exploratory Text Analytic System. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '10)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/1835804.1835827

[129] Stephan Weibelzahl, Alexandros Paramythis, and Judith Masthoff. 2020. Evaluation of Adaptive Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 394–395. https://doi.org/10.1145/3340631.3398668

[130] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[131] 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

[132] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[133] Meng Xia, Mingfei Sun, Huan Wei, Qing Chen, Yong Wang, Lei Shi, Huamin Qu, and Xiaojuan Ma. 2019. PeerLens: Peer-inspired Interactive Learning Path Planning in Online Question Pool. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019-05-02) *(CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300864

[134] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A Visual Interactive System to Enhance the Persuasiveness of Arguments in Online Discussion. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 319 (nov 2022), 30 pages. https://doi.org/10.1145/3555210

[135] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[136] Sha Yuan, Yu Zhang, Jie Tang, Wendy Hall, and Juan Bautista Cabotà. 2020. Expert finding in community question answering: a review. *Artificial Intelligence Review* 53, 2 (2020), 843–874.

[137] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

[138] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.

[139] Zhi-Min Zhou, Man Lan, Zheng-Yu Niu, and Yue Lu. 2012. Exploiting user profile information for answer ranking in cqa. In *Proceedings of the 21st international conference on World Wide Web*. 767–774.