# Creative Research Question Generation for Human-Computer Interaction Research

Yiren Liu[1,†], Mengxia Yu[2,†], Meng Jiang[2] and Yun Huang[2]

[1]University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA
[2]University of Notre Dame, Notre Dame, IN, 46556, USA

**Abstract**

It is essential to develop innovative and original research questions/ideas for interdisciplinary research fields, such as Human-Computer Interaction (HCI). In this work, we focus on discussing how recent natural language generation (NLG) methodologies can be applied to promote the formulation of creative research questions. We collect and curate a dataset that contains texts of RQs and related work sections from HCI papers, and introduce a new NLG task of automatic HCI research question (RQ) generation. In addition to applying common NLG metrics used to evaluate generation accuracy, including ROUGE and BERTScore, we propose two sets of new metrics for evaluating the creativity of generated RQs: 1) DistGain and DiffBS for novelty, and 2) PPLGain for the level of surprise. The task is challenging due to the lack of external knowledge. We investigate four approaches to enhance the generation models with (1) general world knowledge, (2) task knowledge, (3) transferred knowledge, and (4) retrieved knowledge. The results of the experiment indicate that the incorporation of additional knowledge benefits both the accuracy and creativity of RQ generation. The dataset used in this study can be found at: https://github.com/yiren-liu/HAI-GEN-release.

**Keywords**

datasets, text generation, creativity

## 1. Introduction

Asking novel research questions (RQ) is key to starting innovative scientific studies. As David Hilbert states, "*he who seeks for methods without having an infinite problem in mind seeks for the most part in 'vain'*". Proficient scientists read and analyze representative literature in a specific domain, in order to identify the limitations of the existing work and ask new RQs [1]. In computer science research, methodologies are often derived from a study's core research question(s) [1]. Research questions (RQs) are one of the most important components in HCI research, which are often explicitly stated in research papers from the HCI domain. As an outline of the whole paper, RQs are often proposed at the beginning sections and often stated in a unified format, e.g., "RQ1: ..., RQ2: ...". For example, in the HCI paper from Lee et al. [2], the authors listed two RQs at the end of the related work section:

> *"RQ1: How do different chatting styles influence people's self-disclosure? and RQ2: How do different chatting styles influence people's self-disclosure over time?"*

Besides the important role of RQs, the interdisciplinary nature of HCI research motivates us to perform this study [3, 4]. There is a global trend of interdisciplinary research [5, 6]. The fact that HCI is a highly interdisciplinary field [3, 4] poses unique challenges [7, 8] to education and research. Depending on their interests and skills, students and scholars could conduct HCI works and contribute from various perspectives to different disciplinary areas [9]. Because of the interdisciplinary nature, HCI research could make contributions that are technical-driven, UX-focused, and/or method-oriented, etc. [10], which opens a wide door to innovation. The related work section of HCI papers often reviews the relevant literature, which can be used to drive the RQs.

Human researchers have to spend a lot of time reading and understanding tons of interdisciplinary literature. Artificial intelligence (AI) systems have demonstrated their abilities to facilitate some types of scientific research tasks, such as summarizing scientific literature [11], recommending related work [12], and generating new biomedical hypotheses [13]. If a machine could generate RQs based on existing literature, it would help HCI researchers discover potential research topics, though they needed to verify the machine-suggested RQ candidates. However, to the best of our knowledge, there is no AI model or research on automating HCI RQ generation. search on automating HCI RQ generation.

In this work, we propose a novel task of research question generation in the field of HCI research. Given the related work section (denoted as *RelatedWork*), the task aims to generate one or multiple research questions. We

*Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia*

[†]These authors contributed equally.

yirenl2@illinois.edu (Y. Liu); myu2@nd.edu (M. Yu); mjiang2@nd.edu (M. Jiang); yunhuang@illinois.edu (Y. Huang)
0000-0003-1507-0303 (Y. Liu); 0000-0002-6627-2709 (M. Yu); 0000-0002-3009-519X (M. Jiang); 0000-0003-0399-8032 (Y. Huang)

notice that given a set of literature, it is easy to come up with plausible but too generic RQs on broad research topics. Therefore, the challenge of our task lies in that when the same set of literature is surveyed from different perspectives, i.e., given different *RelatedWork*, the generated RQs should be different, correspondingly.

To study this new problem, we build a dataset from HCI literature. We collect 8,904 HCI papers from *Arxiv* and manually extract 158 data examples. Each example has the text of the related work section and the text of research questions. In this study, we develop and evaluate four approaches: (1) prompting pre-trained GPT-3 [14] that has knowledge from pre-training corpus, (2) BART [15] that is fine-tuned on our limited training examples, (3) transfer learning for knowledge augmentation that warms up the model to generate paper titles which are much more accessible than RQs, and (4) retrieval-based augmentation that uses information from the HCI literature text we provide.

We evaluate the RQ generation quality based on three sets of automated metrics: (1) ROUGE and BERTScore with target RQs as references for *accuracy*, (2) DistGain and DiffBS for *novelty*, and (3) PPLGain for *level of surprise*. We propose to use these metrics for evaluation for practical reasons. First, when RQs are not explicitly spelled out in HCI papers, the model that yields greater accuracy could be more effectively utilized. As researchers try to quickly form the RQs given a large amount of surveyed papers, the model could aid in boosting the efficiency of literature review for both research and learning purposes. Second, the model that leads to higher novelty and surprises could be used, when the HCI papers already explicitly present RQs. In this case, researchers can compare the existing "ground truth" RQs with the generated RQs to explore "new" directions for future research.

The main contributions of this study are:

- We propose the task of HCI research question generation, collecting and releasing a dataset.
- We design and develop four types of models that leverage various knowledge to improve RQ generation.
- We evaluate the accuracy, novelty, and level of surprise of generated RQs and find that knowledge transfer is the most promising approach when the available task data size is small.

# 2. Related Work

## 2.1. Question Generation

Automatic question generation (QG) has been studied as a data augmentation approach for Question Answering [16] and Machine Reading Comprehension [17]. Most existing QG studies focus on factoid questions, whose answers are short pieces of text. The QG datasets are usually converted from the question-answering datasets. Instead of factoid questions, RQs are open-ended questions, the generation of which is found to be more challenging in prior work [18], because it requires a deep understanding and needs to be addressed with long-form answers. Nevertheless, the existing open-ended question generation tasks are conditioned on the answers. More research is needed to be done in order to generate unsolved open-ended HCI research problems.

For educational domains, QG systems often aim at generating assessment questions, e.g., multi-choice questions, to help students understand the learning materials and reduce the manual workload required from instructors. Emerging studies have proposed datasets for educational QG [19, 20, 21]. However, these works aim to generate questions that help with comprehension of learning materials, not exploring potential unsolved research problems.

## 2.2. Scientific Text Generation

In order to reduce the burden of scientific writing or simulate scientists' behaviors, there is a line of research aiming at automatic scientific text generation. Since early work on abstract generation [22], various approaches have been proposed for scientific text summarization [23, 24, 11]. Spangler et al. [13] leverage text mining for scientific hypothesis generation. ReviewerBot [25] utilizes information extracted from knowledge graphs to construct synthetic paper reviews from templates. AutoCite [26] leverages multi-modal information to generate contextualized citation texts. PaperRobot [27] cascadingly generates abstracts, conclusions, future work, and titles for a follow-on paper. However, the automatic HCI RQ question has not been studied as an NLP task.

## 2.3. Evaluating Creativity in Text Generation

Methods for enhancing the ability of machine learning models to produce original content have been a crucial topic in the emerging research domain of computational creativity [28]. Franceschelli and Musolesi [29] summarized existing methods for creativity evaluation and discussed their potential application in recent deep learning models (e.g., VAE and GAN). However, most of these existing evaluation methods are highly subjective and require strong human intervention. With the recent advances in text generation methods based on pre-trained language models, additional research is still needed to be done in order to automatically and objectively evaluate the creativity of text generation models. Prior NLP research has discussed potential methods to automatically evaluate generation taking into consideration both

| | Avg. # of words per Related Work | per RQ | Avg. # of RQs per paper |
|---|---|---|---|
| **train** | 409.9 | 14.2 | 2.4 |
| **dev** | 369.9 | 13.8 | 2.4 |
| **test** | 332.4 | 15.7 | 2.2 |

**Table 1**
Descriptive Statistics of the Proposed Dataset

the accuracy and diversity of the generated results [30]. In this work, we employ Boden's three criteria [31] for studying machine creativity, defined as *"the ability to generate ideas or artifacts that are new, surprising and valuable"*, to propose new metrics for creativity evaluate in text generation tasks.

## 3. Problem Definition and Data

A research question refers to a question that a study or research project aims to address. In HCI research publications, RQs are often proposed after the survey of related work. Based on the understanding of existing literature and citation purposes, different papers will compose the related work sections differently, even if they cite the same set of literature. Correspondingly, their research questions should be different.

We formally define the task of HCI RQ generation with task variables as follows.

**Definition 1 (HCI Research Question Generation).** *Given the RelatedWork of an HCI research paper, the generation model requires maximizing* $P(RQ|RelatedWork)$.

In real-life HCI research scenarios, researchers strive to propose highly novel and creative research questions based on existing work. Thus, we propose also to measure the creativity of generated research questions. Based on the theory of Boden's criteria [31] *"the ability to generate ideas or artifacts that are new, surprising, and valuable"*, we construct the creativity measurement as a combination of two aspects: 1) *novelty* and 2) *level of surprise*. We do not evaluate the *value* of generated RQs since we believe it would require extensive expert knowledge and is hardly feasible without human intervention.

**Definition 2 (Generation Creativity).** *1) We measure the* ***novelty*** *of a set of RQs by comparing their similarity to the RQs of prior publications within our collected corpus; 2) We measure the* ***level of surprise*** *of a set of RQs based on their perplexity with respect to the perplexity of existing RQs using a large PLM (e.g., GPT-2).*

To collect open-access HCI publications, we used papers available through *Arxiv*. We collected PDF files of papers under the category of Human-Computer Interaction (cs.HC) [1] using the public API provided by *Arxiv*. This resulted in a total of 8,904 HCI-related papers [2]. We then convert these papers from PDF to sectioned XML format using *GROBID*[3] and *SciPDF Parser*[4] in order to further analyze and filter based on their textual context. The section and title information are preserved in the XML version of our collected papers. For research questions, we conducted pattern matching of question sentences starting with "RQ". In order to collect text from related work sections, i.e., *RelatedWork*, we extract sections with titles containing the keywords "related work". We remove RQs from *RelatedWork* if it appears.

The resulting dataset consists of 158 valid examples. We then split the dataset into train/dev/test sets with 108/25/25 examples. Note that the splits are carefully arranged in chronological order, i.e. papers in the dev and set are published later than those from the train split. This is to ensure the RQs in the dev/test sets are the newest and are not revealed in the train set. The descriptive statistics of the final dataset can be found in Table 1.

## 4. Method

To tackle the lack of knowledge issue in HCI RQ generation, we investigate four types of approaches that leverage different types of knowledge. We present three sets of quantitative metrics to evaluate the quality of generated questions from three different aspects: accuracy, novelty and level of surprise.

### 4.1. Generation Models with Various Knowledge

In this section, we describe the different models used for training and evaluating the RQ generation task.

**Pre-trained GPT-3** As a large language model (LM) with 175 billion parameters, GPT-3 is the state-of-the-art learner succeeding on many NLP tasks and shows its capability in research paper writing [32], educational question generation [33] and open-domain QA [34]. GPT-3 is trained on 45 TB of text data from multiple sources which include Wikipedia and books, enabling the model to store a huge amount of **general world knowledge**.

**Fine-tuned BART** We choose BART, a Transformers-based pretrained generation model, as our backbone model. By fine-tuning BART on our RQ generation dataset, the model should acquire specific **task knowledge**, but the knowledge would be limited due to data scarcity.

**Knowledge transfer from title generation** Transfer learning is an effective way to improve the model

[1] https://arxiv.org/list/cs.HC/recent

[2] https://github.com/yiren-liu/HAI-GEN-release

[3] https://github.com/kermitt2/grobid

[4] https://github.com/titipata/scipdf_parser

when only a limited amount of data on the target task is available. The available RQ data may be limited for a variety of reasons, e.g., errors during PDF parsing, or RQs that are not explicitly written in some papers. In contrast, paper titles are more accessible, where the amount we extracted is 30 times that of research questions. In semantic space, a paper's title represents its most significant contribution, which is strongly tied to its research topics. In most cases, paper titles can be considered as a high-level summary of the solution to the research questions. Therefore, we propose to augment the BART model with transfer **relevant task knowledge** from title generation to RQ generation. The titles in train/dev/test sets are excluded. They are not used as input for the target task. So there is no data leaking.

**Knowledge retrieval from HCI corpus** Knowledge retrieval is another promising solution to many knowledge-intensive NLP tasks [35] such as question answering [36] and information-seeking question generation [37]. To incorporate **external domain knowledge**, we apply the Dense Passage Retriever (DPR) [36] to retrieve sentences most relevant to the input *RelatedWork* from the HCI corpus. The retrieved sentences are appended to the end of the original related work text as input.

## 4.2. Evaluation Methods for Novelty and Surprise

The task of HCI RQ generation aims to generate open-ended research questions to inspire researchers, which need to be highly creative. Recently, Computational Creativity has become an emerging field of study in the HCI domain [29]. Inspired by Boden's three criteria [31] *"the ability to generate ideas or artifacts that are new, surprising and valuable"*, we introduce evaluation metrics to measure the *novelty* and *level of surprise* of generated RQs. We do not evaluate the *value* of generated RQs to HCI research since it would require extensive expert knowledge and human intervention.

### 4.2.1. Measuring novelty

To evaluate the novelty of generated RQs, i.e., how new/original the RQs are, we measure the difference between the generated RQs and prior RQs. We introduce two metrics: 1) an *n*-gram-based score **DistGain**, and an embedding-based score **DiffBS**. We first make a set of prior RQs, denoted as $\{RQ_{\text{existing}}\}$, from papers published earlier than the papers in dev/test sets.

**Distinct-*k* gain (DistGain** or **DG)** is defined based on Distinct-*k* [38]. We calculate the average proportion of new unique *n*-grams in the newly generated RQ compared to the total number of *n*-grams in the $\{RQ_{\text{existing}}\}$. DistGain can be written as follows:

$$DistGain_j = \frac{1}{M}\sum_{i=1}^{M}\frac{|\{y_j\}-\{x_i\}|}{|Y_j|}, \tag{1}$$

where sequence $Y_j = (y_j)_{:}$ denotes the *j*-th generated RQ, sequence $X_i = (x_i)_{:}$ denotes the *i*-th RQ in $\{RQ_{\text{existing}}\}$, and $M = |\{RQ_{\text{existing}}\}|$. We average the $DistGain_j$ of all generated RQs to obtain an overall score *DistGain*.

**Difference in BERTScore (DiffBS** or **DBS)**: In order to measure the distance between the generated RQ and $\{RQ_{\text{existing}}\}$, we calculate cosine similarity of BERT embeddings [39] between the generated RQ and each $X_i \in \{RQ_{\text{existing}}\}$. For each generated RQ, we calculate the F1-BERTScore for each pair $(Y_j, X_i)$, and average over all existing RQs:

$$DiffBS_j = \frac{1}{M}\sum_{i=1}^{M}(1 - F_{\text{BERT}}(Y_j, X_i)), \tag{2}$$

where $F_{\text{BERT}}(Y_j, X_i)$ denotes the F1-BERTScore calculated between $Y_j$ and $X_i$. The final $DiffBS$ for each model is averaged over all generated RQs.

### 4.2.2. Measuring level of surprise

To measure the level of surprise, we refer to Boden [31]'s definition of surprise *"an idea may be surprising because it's unfamiliar, or even unlikely"*. We propose a new automatic metric to measure the level of surprise in generated RQs.

**Perplexity Gain (PPLGain)**. Perplexity, the inverse probability, is frequently used to measure how uncertain an LM generates the test data. Given a text, the higher the perplexity is, the more uncertain the LM is about generating it. Assuming an LM is successfully pre-trained with a sufficient amount of general text data, the perplexity reflects the unexpectedness, or level of surprise, of the LM to the given text. Thus, we employ the perplexity of GPT-2 of the RQs:

$$ppl(Y_j) = \exp\left(-\frac{1}{T}\sum_{i=1}^{T}\log\, p(y_i|y_1, ..., y_{i-1})\right). \tag{3}$$

To measure the level of surprise, or unexpectedness, of the generated RQs, we calculate the difference between the perplexity of generated RQs and prior RQs. We define the perplexity gain as follows:

$$PPLGain_j = \frac{ppl(Y_j) - \frac{1}{M}\sum_{i=1}^{M} ppl(X_i)}{\frac{1}{M}\sum_{i=1}^{M} ppl(X_i)}. \tag{4}$$

The final *PPLGain* score is averaged over all $Y_j$.

# 5. Experiments

## 5.1. Evaluation Methods

We evaluate the generation quality with three sets of metrics: (1) ROUGE and BERTScore for measuring accuracy; (2) DistGain and DiffBS for measuring novelty; (3) PPLGain for measuring surprise.

## 5.2. Experimental Settings

We evaluated four text generation models with different types of knowledge over our proposed dataset.

**GPT-3**. We prompt GPT-3 (text-davinci-002) with a one-shot example. We use a temperature of 0.7 and pick the top-1 generation. To align the output format with BART-based models, we post-process the GPT-3 output by replacing the question number. That means, "1." or "1)" will be replaced by "RQ1:".

**BART-FT**. We use the *RelatedWork* section as input. An HCI paper may have multiple research questions. The latter ones are highly likely to be dependent on the previous ones. Thus, instead of an individual RQ, our output is set as a sequence of concatenated RQs such as "RQ1: ..., RQ2: ...". For all the experiments with the BART model, the maximum input and output length is set as 768 and 128 tokens, respectively.

**BART-FT+transfer**. To transfer knowledge from title generation, we first fine-tune the BART model on {*RelatedWork*, title} pairs and then continue fine-tuning on {*RelatedWork*, RQ} pairs. We carefully construct the dataset for title generation and avoid dev/test RQ data leaking in the training data of the title generation.

**BART-FT+retrieval**. To construct the retrieval corpus, we gather the abstract, introduction, and related work section of the existing papers that were published before dev/test papers, split the text into sentences, and form an HCI corpus containing 310,955 sentences. We retrieve top-3 sentences with pre-trained DPR using related work as queries, and append the retrieved text to input sequences.

## 5.3. Results

Results on automatic evaluation are presented in Table 2.

***GPT-3 with general world knowledge increases generation novelty, but under-performs fine-tuned models in accuracy and surprise level***. Table 2 shows that, compared to the BART models, GPT-3 performs worse in terms of ROUGE and BERTScore on dev and test, but it surpassed the other three models on DistGain and DiffBS, which are both measurements for generation novelty. However, all three BART-based models achieved higher PPLGain scores which measure the level of surprise of generated RQs. As a large LM, GPT-3 possesses rich knowledge outside of the HCI research domain, which enables it to output different words from existing RQs, but those words may be off the research topic.

***Knowledge augmentation is effective on HCI RQ generation***. Transfer learning augmented model, i.e., BART-FT+transfer, outperforms BART baselines in terms of ROUGE-2 (11.7%↑ on dev and 9.1%↑ on test) and ROUGE-L (4.7%↑ on dev and 3.1%↑ on test). The effectiveness of transfer learning shows that learning the task of title generation helps bridge the gap between existing research and new research. Retrieval augmented models, i.e., BART-FT+retrieval, perform at the same level as BART-FT on dev set and significantly outperforms BART-FT in terms of ROUGE-2 (20.9%↑) and ROUGE-L (4.2%↑) on the test set. The model also surpassed the baseline BART-FT in novelty and surprise metrics. Both knowledge augmentation methods improve the novelty and surprise of the generated RQs. This implies that introducing additional knowledge from publications enables the language model to generate RQs with new ideas outside the training set. Although both methods improve generation novelty and surprise, using knowledge transfer results in a higher increase. This might be because titles tend to reflect the contributions of studies in a self-contained and abstractive manner. Similarity-based retrieved results tend to be individual sentences that might be confusing, or even noisy when they are used as input, because they bring information outside the context paragraph.

# 6. Discussion

## 6.1. Case Study of Generated RQs

To further validate the proposed creativity metrics, we qualitatively compare examples of RQs generated by different models, as shown in Table. 3. It shows that RQs generated by GPT-3 appear to be less relevant compared to other models, where the research topic is generalized from "GitHub issues" to "online discussion". Meanwhile, the results generated by GPT-3 also suffered from repetition as the sequence of "incivility and toxicity in online discussions" appeared twice in the given example. However, the language/words it uses could be new compared to prior RQs. This implies that the incorporation of *general world knowledge* generalizes the content of machine-created RQs to domains other than that of the target paper. In this example, only BART-FT+transfer captured the information about "maintainers" which is critical in the ground truth RQ2, showing the advantage of transfer learning. We also found that the output of BART achieved the highest PPLGain score (level of surprise), as the results mentioned interesting concepts including

| | Dev | | | | | | Test | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **R-2** | **R-L** | **BS** | **DG** | **DBS** | **PG** | **R-2** | **R-L** | **BS** | **DG** | **DBS** | **PG** |
| **GPT-3** | 8.48 | 21.1 | 80.26 | **78.1** | **13.1** | -74.9 | 10.27 | 22.97 | 80.42 | **76.9** | **13.2** | -72.2 |
| **BART-FT** | 12.92$^{***}$ | 26.88$^{***}$ | 83.48$^{***}$ | 60.1 | 10.4 | -45.9$^{***}$ | 12.37 | 26.65 | 82.57$^{**}$ | 59.9 | 10.5 | -39.9$^{***}$ |
| **BART-FT+transfer** | **14.43**$^{***}$ | **28.14**$^{***}$ | **84.04**$^{***}$ | 65.1 | 10.8 | **-37.1**$^{***}$ | 13.5 | 27.48$^{*}$ | **82.97**$^{**}$ | 64.5 | 11.1 | **-27.6**$^{***}$ |
| **BART-FT+retrieval** | 12.87$^{**}$ | 26.93$^{**}$ | 83.22$^{***}$ | 60.5 | 10.5 | -42.1$^{***}$ | **14.96** | **27.76**$^{*}$ | 82.54$^{***}$ | 63.2 | 10.8 | -35.5$^{***}$ |

* Metric notations: R-2: **ROUGE-2**, R-L: **ROUGE-L**, BS: **BERTSore**, DG: **DistGain**, DBS: **DiffBS**, PG: **PPLGain**. For all metrics, higher means better. T-tests are conducted for all BART-based models using GPT-3 as a baseline (with $p < 0.001^{***}, < 0.01^{**}, < 0.05^{*}$).

**Table 2**
Automatic evaluation results of four models on HCI RQ generation. **GPT-3** is prompted with one-shot examples. **BART-FT** denotes BART fine-tuned on our dataset. **BART-FT+transfer** denotes BART fine-tuned with transfer learning. **BART-FT+retrieval** denotes BART fine-tuned with knowledge retrieval.

| **Gold** | **RQ1: What are the characteristics of github locked issues? RQ2: What are the justifications given in the comments by project maintainers when locking issues as too heated?** | **R-2** | **R-L** | **BS** | **DG** | **DBS** | **PG** |
|---|---|---|---|---|---|---|---|
| **GPT-3** | RQ1: How effective are github locked issues in curbing incivility and toxicity in online discussions? RQ2: How well do current machine learning detectors perform in detecting incivility and toxicity in online discussions? | 7.7 | 23.7 | 81.2 | 86.1 | 13.4 | -79.7 |
| **BART** | RQ1: What kinds of incivility are present in code review discussions of rejected patches? RQ2: How common are code review issues locked as too heated? | 11.5 | 38.4 | 86.7 | 69.7 | **13.8** | **-46.5** |
| **BART-FT+transfer** | RQ1: How do maintainers respond to github issues locked as too heated? RQ2: What are maintainers' reactions to the locked issues? | 19.2 | 33.3 | **90.3** | 82.4 | 13.4 | -53.9 |
| **BART-FT+retrieval** | RQ1: What kinds of incivility exist in github issues locked as too heated? RQ2: What are the most common types of incivility in github? | **23.0** | **39.2** | 87.5 | **90.9** | 13.5 | -59.9 |

**Table 3**
Generated RQs on a test example of the paper titled *"How heated is it? Understanding GitHub locked issues"*.

"code review" and "rejected patches".

## 6.2. Limitations and Future Work

Although the experimental results revealed RQ generation as a promising and meaningful task, several limitations exist in our current study. First, the training and evaluation of generation methods were conducted on a relatively small-scale dataset, undermining the solidity of the conclusions yielded from the experiments. Future work should consider expanding the dataset by collecting more open-access publications and employing careful human annotation to expand the scale and improve the quality of the dataset. Second, the evaluation metrics used/proposed in this work did not fully consider the open-ended nature of the RQ generation tasks. In practice, a well-surveyed research topic should yield many open-ended creative research questions, while our evaluation was sorely based on the comparison between the generated and ground-truth RQs. Further quantifiable human evaluation should be incorporated to validate the quality of generation. Additionally, the evaluation of GPT-3 as an RQ generation method only covered a one-shot scenario with a manually selected example by researchers. Future work should take into consideration the potential impact of the demonstration selection method on the generation quality of GPT-3.

# 7. Conclusions

In this work, we proposed a novel NLP task of HCI RQ generation. We curated a dataset of 8,904 HCI publications and a collection of 158 examples of (related work, RQ)-pairs. In addition to accuracy metrics, we evaluated the creativity of RQ generation with metrics for novelty and surprise. We investigated the performance of four approaches that leverage different types of knowledge. Through experiments, we showed general world knowledge in pre-trained LM helped improve generation novelty, and domain knowledge augmentation methods improved accuracy and level of surprise. Future studies could explore knowledge augmentation methods by incorporating different kinds of knowledge, e.g., general world knowledge, task knowledge, transferred domain knowledge, or retrieved textual knowledge.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2119589. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

[1] R. Elio, J. Hoover, I. Nikolaidis, M. Salavatipour, L. Stewart, K. Wong, About computing science research methodology, 2011.

[2] Y.-C. Lee, N. Yamashita, Y. Huang, W. Fu, " i hear you, i feel you": encouraging deep self-disclosure through a chatbot, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–12.

[3] H. R. Hartson, Human–computer interaction: Interdisciplinary roots and trends, Journal of systems and software 43 (1998) 103–118.

[4] J. Bardzell, S. Bardzell, Humanistic hci, Synthesis Lectures on Human-Centered Informatics 8 (2015) 1–185.

[5] C. Kim, H. Kim, S. H. Han, C. Kim, M. K. Kim, S. H. Park, Developing a technology roadmap for construction r&d through interdisciplinary research efforts, Automation in Construction 18 (2009) 330–337.

[6] D. Rhoten, Interdisciplinary research: Trend or transition, Items and Issues 5 (2004) 6–11.

[7] C. Rusu, V. Rusu, Teaching hci: a challenging intercultural, interdisciplinary, cross-field experience, in: International Workshop on Intercultural Collaboration, Springer, 2007, pp. 344–354.

[8] P. Dourish, J. Finlay, P. Sengers, P. Wright, Reflective hci: Towards a critical technical practice, in: CHI'04 extended abstracts on Human factors in computing systems, 2004, pp. 1727–1728.

[9] W. E. Mackay, A.-L. Fayard, Hci, natural science and design: a framework for triangulation across disciplines, in: Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques, 1997, pp. 223–234.

[10] J. Lazar, J. H. Feng, H. Hochheiser, Research methods in human-computer interaction, Morgan Kaufmann, 2017.

[11] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, D. R. Radev, Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 7386–7393.

[12] M. Färber, A. Jatowt, Citation recommendation: approaches and datasets, International Journal on Digital Libraries 21 (2020) 375–405.

[13] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers, et al., Automated hypothesis generation based on mining scientific literature, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 1877–1886.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, 2020, pp. 7871–7880.

[16] N. Duan, D. Tang, P. Chen, M. Zhou, Question generation for question answering, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 866–874.

[17] R. Puri, R. Spring, M. Shoeybi, M. Patwary, B. Catanzaro, Training question answering models from synthetic data, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5811–5826. URL: https://aclanthology.org/2020.emnlp-main.468. doi:10.18653/v1/2020.emnlp-main.468.

[18] S. Cao, L. Wang, Controllable open-ended question generation with a new question type ontology, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6424–6439.

[19] G. Chen, J. Yang, C. Hauff, G.-J. Houben, Learningq: a large-scale dataset for educational question generation, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[20] H. Gong, L. Pan, H. Hu, Khanq: A dataset for generating deep questions in education, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 5925–5938.

[21] S. Pollak, V. Podpecan, J. Kranjc, B. Lesjak, N. Lavrac, Scientific question generation: Pattern-based and graph-based robochair methods., in: ICCC, 2021, pp. 140–148.

[22] K. Ono, K. Sumita, S. M. Research, D. Center, T. C. Komukai-Toshiba-cho, et al., Abstract generation

based on rhetorical structure extraction, arXiv preprint cmp-lg/9411023 (1994).
[23] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, arXiv preprint arXiv:1804.05685 (2018).
[24] I. Cachola, K. Lo, A. Cohan, D. S. Weld, Tldr: Extreme summarization of scientific documents, arXiv preprint arXiv:2004.15011 (2020).
[25] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, N. F. Rajani, Reviewrobot: Explainable paper review generation based on knowledge synthesis, arXiv preprint arXiv:2010.06119 (2020).
[26] Q. Wang, Y. Xiong, Y. Zhang, J. Zhang, Y. Zhu, Autocite: Multi-modal representation fusion for contextual citation generation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 788–796.
[27] Q. Wang, L. Huang, Z. Jiang, K. Knight, H. Ji, M. Bansal, Y. Luan, Paperrobot: Incremental draft generation of scientific ideas, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1980–1991.
[28] A. Cardoso, T. Veale, G. A. Wiggins, Converging on the divergent: The history (and future) of the international joint workshops in computational creativity, AI magazine 30 (2009) 15–15.
[29] G. Franceschelli, M. Musolesi, Creativity and machine learning: A survey, arXiv preprint arXiv:2104.02726 (2021).
[30] W. Yu, C. Zhu, T. Zhao, Z. Guo, M. Jiang, Sentence-permuted paragraph generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5051–5062.
[31] M. A. Boden, The creative mind: Myths and mechanisms, Routledge, 2004.
[32] M. Lee, P. Liang, Q. Yang, Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities, in: CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–19.
[33] S. Wang, Y. Liu, Y. Xu, C. Zhu, M. Zeng, Want to reduce labeling cost? gpt-3 can help, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4195–4205.
[34] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, M. Jiang, Generate rather than retrieve: Large language models are strong context generators, in: International Conference for Learning Representation (ICLR), 2023.
[35] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.
[36] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.
[37] M. Gaur, K. Gunaratna, V. Srinivasan, H. Jin, Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 10672–10680.
[38] J. Li, M. Galley, C. Brockett, J. Gao, W. B. Dolan, A diversity-promoting objective function for neural conversation models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 110–119.
[39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).