Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion

Hyunseung Lim*
KAIST
Daejeon, Republic of Korea
charlie9807@kaist.ac.kr

Ji Yong Cho*
LG AI Research
Cornell University
Seoul, Republic of Korea
jiyong.cho@lgresearch.ai

Taewan Kim
KAIST
Daejeon, Republic of Korea
taewan@kaist.ac.kr

Jeongeon Park†
KAIST
Daejeon, Republic of Korea
jeongeon.park@kaist.ac.kr

Hyungyu Shin
KAIST
Daejeon, Republic of Korea
hyungyu.sh@kaist.ac.kr

Seulgi Choi
KAIST
Daejeon, Republic of Korea
igules8925@kaist.ac.kr

Sunghyun Park
LG AI Research
Seoul, Republic of Korea
sunghyun.park@lgresearch.ai

Kyungjae Lee
LG AI Research
Seoul, Republic of Korea
kyungjae.lee@lgresearch.ai

Juho Kim
KAIST
Daejeon, Republic of Korea
juhokim@kaist.ac.kr

Moontae Lee
LG AI Research
University of Illinois Chicago
Seoul, Republic of Korea
moontae.lee@lgresearch.ai

Hwajung Hong
KAIST
Daejeon, Republic of Korea
hwajung@kaist.ac.kr

ABSTRACT

Research promotion enables researchers to share advanced knowledge with pertinent academic communities. The question-and-answer (QA) style articles are effective for researchers to promote their research by enabling readers to understand research on complex subjects. Recent advances in large language models (LLMs) have opened avenues for supporting researchers in creating QA-style articles for research promotion. However, without the authors' involvement, these models may only partially capture the researcher's intention and voice. We developed AQUA, a research probe that enables researchers to co-create QA-style articles with LLMs to promote their research papers. A user study (n=12) reveals that LLMs reduced authors' burden and helped them understand the readers' perspectives. Nevertheless, LLMs failed to capture the unique intent of the authors, and their automated generation discouraged authors from carefully revising their answers. Based on our findings, we discuss human-LLM interaction design to enable authors to create QA-style articles that reflect their intention.

CCS CONCEPTS

- **Human-centered computing → Human computer interaction (HCI)**; **Interaction design**.

KEYWORDS

Large Language Model, Human-AI Interaction, Question-and-Answer, Research Promotion

ACM Reference Format:

Hyunseung Lim, Ji Yong Cho, Taewan Kim, Jeongeon Park, Hyungyu Shin, Seulgi Choi, Sunghyun Park, Kyungjae Lee, Juho Kim, Moontae Lee, and Hwajung Hong. 2024. Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3643834.3660705

1 INTRODUCTION

Research promotion is crucial for researchers in communicating advanced knowledge with communities with similar academic interests. It is also important to disseminate research findings across a spectrum of audiences beyond the confines of specific academic fields the research is based on for fostering interdisciplinary collaboration and driving innovation in science [24, 30, 69, 74] and for enhancing the visibility of researchers and establishing researcher identities as experts [1, 12, 25]. However, writing for research promotion targeting a wide range of audiences is challenging as effective promotional content entails both scientific writing techniques to deliver scientific knowledge found from research and creative writing techniques to translate highly specialized information into

---

[*] Authors contributed equally.

[†] Jeongeon's current affiliation is at DGIST, South Korea.

This work is licensed under a Creative Commons Attribution International 4.0 License.

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3660705

a digestible narrative, attract readers, and draw them into the narrative [25].

Writing promotional content in a question-and-answer (QA) style can be a viable solution. In education, teaching and learning through asking and answering questions is considered an effective pedagogical approach, especially for complex subjects [2, 27, 55, 73]. Similarly, research papers, which deliver complex ideas, benefit from the QA-style promotion; questions unpack an intricate concept and act as a roadmap to building the necessary elements of knowledge for readers to comprehend the paper's scientific findings. Many science podcasts, pre-recorded interviews with experts, exemplify this approach [e.g., 44, 54, 57]. Hosts ask questions one by one to understand research findings on behalf of a broad audience with insufficient expert knowledge. Even in research papers, authors include frequently asked questions (FAQs) to provide additional support for readers to understand their research [e.g., 15, 67].

Recent advances in large language models (LLMs) have opened up opportunities to further support researchers with writing. Prior research has shown LLMs can reduce writers' cognitive load and brainstorming ideas by generating drafts for both scientific writing [e.g., 23, 28] and argumentative writing [77]. Nevertheless, fully automating draft generation without involving writers is less likely to fulfill writers' intentions [41, 43]. Therefore, incorporating human engagement throughout the writing process is considered the key design imperative to produce desired writing outcomes [14, 16, 59, 77]. The present study explores how researchers interact with LLMs in writing QA-style articles, focusing on researchers' behavioral engagement in the writing process, their perceptions of LLMs' writing support, and the interplay between the two.

We developed Authors' Question-and-Answers (AQUA), a research probe that enables researchers to co-create QA-style articles with LLMs to promote their research papers. AQUA recommends different types of questions—general, personalized, and follow-up questions—and drafts answers for users, and users build a narrative in a QA-style article by selecting and ordering these question-and-answer pairs as building blocks. Using AQUA as a probe, the present study aims to understand interactions between users and LLMs in the co-creation process and derive design implications for human-AI co-creation systems. The contributions of our paper are as follows:

- A probe study to examine how researchers write QA-style articles with LLMs for promoting research papers.
- Findings of our probe study on how the authors interacted with and perceived LLM-powered features of AQUA.
- Design implications for human-LLM interaction in co-creating QA-style articles for research promotion.

2 RELATED WORK

2.1 Benefits of Question-and-Answer (QA) Style Promotion

Asking questions and answering them has long been an effective way of attaining knowledge. Theoretically, questions stimulate curiosity as well as high-order thinking, such as critical thinking [73]. Finding answers to a series of questions allows learners to connect ideas and discover new ideas. For example, the Socratic method is a pedagogical approach still being incorporated at different levels of education [50]. It originated from the Greek philosopher Socrates, who taught his pupils with continuous questions. A more modern pedagogical approach that highly values QAs is inquiry-based learning [2]. It posits that students can best learn when they engage with learning materials (e.g., textbooks, experiments) to answer an interesting question (e.g., finding references and collecting evidence to answer the questions) [27]. While teachers are the ones who ask questions using the Socratic method, inquiry-based learning emphasizes students' agency in developing their own questions. Despite the difference, both approaches indicate that the QA style mirrors a knowledge acquisition process.

In addition to effective knowledge acquisition, the QA style can bring joy to the learning process as QAs can be arranged to develop a well-written story. Media shows examples of the QA-style information delivery with a narrative [e.g., 42, 63]. For example, podcasts and talk shows, in which hosts invite guests and interview them, fundamentally consist of QAs, delivering information centering around a topic. Similarly, in written texts, an interview-based article is a standard media piece frequently appearing in magazines. Interview-based articles are also composed of QAs, typically reconstructed after interviewing a person of interest. Media studies have underlined the success of media depends on that narrative, stressing the importance of storytelling to keep audiences' attention and deliver intended information successfully [8].

In sum, the QA-style promotion of a research paper can be an effective format for delivering information as it reflects human learning process. This format can also engage readers with a narrative like a talk show if a series of QA pairs are well structured.

2.2 Supporting QA Generation and Its Applications

In the NLP literature, researchers have tackled tasks regarding generating questions (i.e., *Question Generation*) and answering questions (i.e., *Question Answering*) [7, 32, 60]. Question Generation is a task to make a valid question that is relevant to the provided context, such as documents of a specific domain. Question Answering is a task to generate correct answers, which is often accompanied by extracting relevant pieces of information from given documents. Recently, pre-trained language models, including LLMs like GPT-4 [47], have made technical breakthroughs in the aforementioned tasks, but the remaining key issue is aligning models with humans' values and needs [38, 68].

In HCI research, advanced NLP techniques have been actively explored on AI-embedded support tools for knowledge acquisition [e.g., 5, 13, 34, 41, 61, 66, 71]. Pre-trained language models are high-performing yet accessible to non-experts in NLP; by providing a specific input (i.e., a *prompt*), non-experts in NLP can generate quality output tailored to their needs specified in the prompt [37]. This technique of guiding the model with prompts–*prompting* or *prompt engineering*– enables designing a nuanced, sophisticated tool that empowers users who may not attain the wanted knowledge without the tool [6, 18]. For instance, *Paper Plain* proposed a feature that generates key questions for a medical research paper to help novices comprehend the content of the paper without having heavy medical knowledge beforehand [5]. In addition, LLM-mediated knowledge acquisition applications have become increasingly prevalent [20, 51, 58]; they support researchers in literature research such as writing paper summaries and answering questions about research papers, enabling more active, scalable research communication.

Importantly, finding a balance between utilizing advanced techniques and inviting humans as part of knowledge generation is critical for designing a successful system that produces desired outcomes. Prior research suggests that human interventions are essential to produce high-quality, context-relevant system output. For example, a previous study designed *ReadingQuizMaker*, a system to support teachers in creating quiz questions for students and compared questions generated solely by a language model and those that were collaboratively crafted with teachers. The model-generated questions turned out to be logical but unsuitable for stimulating learning, indicating that teachers' role in the system was essential to create effective questions for educational purposes [41]. Similarly, other studies have found that human-in-the-loop editing is necessary for targeted writing in context, such as writing screenplay scripts that tell a coherent story with well-developed characters and scenes [43].

In sum, advances in NLP techniques provide exciting opportunities to augment human intellectual activities (e.g., reading and writing). Nevertheless, humans still need to get involved in the activity for successful outcomes, which remains a design challenge for AI-embedded support tools.

2.3 Human-AI Co-Creation Systems

Human-AI co-creation systems can be defined as systems in which both humans and AI take part in making an end product (e.g., artifacts, drawings, writings). Admittedly, although the systems could be labeled differently in a more fine-grained manner (e.g., support tools, partners) depending on the leading/supporting role of the system (e.g., humans predominantly lead, AI leads vs. both contribute equally) and the level of engagement of each party (e.g., AI makes comments on humans' work, humans audit AI's work vs. humans and AI collaboratively work), this review focuses on the design implications learned from previous work regardless of labels.

Research on human-AI co-creation systems have been actively conducted in a wide range of creative tasks for design [29], illustrations [9, 62, 65], music composition [40], dancing [64], drawing [19, 33, 46], storytelling [14, 16, 43, 45, 48, 59], and writing [4, 17, 23, 28, 39, 52, 77]. As these creative tasks are characterized as complex and cognitively demanding to humans, AI in the human-AI co-creation systems is designed to facilitate the work process and inspire humans with ideas by taking on arduous work for humans (e.g., searching evidential information for writing [52, 77] and providing various alternatives [28, 43]) and suggesting potential next steps (e.g., proposing a next scene/line of the story [16, 45, 59] and listing potential writing directions [52]). When using these systems, for example, for writing tasks, users engage with the writing activity longer, experience reduced mental load, and become more productive, which they would not have done otherwise without the systems [39, 48, 59, 77]. Particularly, previous research has supported writing for research communication, using NLP-based analysis to inform effective storytelling for specific audiences [4].

Since the introduction of generative AIs, in most human-AI co-creation systems, AI has typically produced work under the supervision of users. For example, users describe what image they want to draw, and then AI draws it for users [e.g., 9], or users provide the purpose of writing, and then AI generates pieces of or the whole writing for users to revise and finalize [e.g., 77]. Interaction design for such systems is central to communicating user intent with AI so that AI delivers what users expect and better assists users during tasks. Previous studies have used different strategies, but they have in common that the system design breaks a complex task into smaller sub-tasks and supports a non-linear and iterative workflow. For example, VISAR [77] scaffolds argument writing by guiding users to set goals in order, setting a scope of context, selecting keywords, selecting discussion points, outlining, and then revising a generated draft. Users can go back and forth between the steps to finalize the draft.

In sum, human-AI co-creation systems flourish in inspiring users with fresh ideas and carrying out laborious work for users in creative tasks, especially after generative AI techniques have become available. Interaction design to communicate user intent with AI is the key to producing end products successfully.

3 FORMATIVE STUDY

We conducted a formative interview study to understand researchers' common practices and challenges in promoting their work and identify opportunities for LLMs to co-create promotional content with researchers as proposed in the previous studies [4, 39, 41]. We recruited early-stage researchers who have led and published their research (authors from now on). As these authors are supposedly more eager to promote their research than more senior and established researchers, interviewing them would produce insights for designing a human-AI co-creation system that facilitates writing promotional content for their research. Consistent with prior studies [4, 39, 41], participants perceived creating promotional content as daunting and arduous; therefore, they shared a positive view of receiving support from LLMs. We derived three design requirements based on the key findings. Our system should help authors (1) understand the perspectives of readers, (2) construct concise yet informative QAs with a narrative, and (3) speak for researcher identities. The details of the formative study are as follows.

3.1 Method

We conducted a semi-structured interview with nine graduate students (four females and five males; one master's student and eight doctoral students). Participants were recruited through an online community affiliated with a large research institute in South Korea. All participants are early career researchers with 2-10 years of research experience, have published research articles in AI/ML and HCI domains, and have promoted their published articles on social media like Twitter. Participants were individually interviewed via Zoom for 40 to 60 minutes. They were asked about (1) their experience of promoting their published research papers on social media platforms such as Twitter, (2) their views on presenting their research in a QA format, and (3) the opportunities and challenges of using the QA format to introduce their work. Three of the authors analyzed interview data using Atlas.ti [3], refining emerging themes iteratively until they reached a consensus.

3.2 Results

3.2.1 DR1: Understanding Perspectives of Readers. All participants highlighted that promoting their articles in the QA style provides unique opportunities to see their work from the readers' perspective. Since the authors are too familiar with their own work, they cannot fathom which parts readers will find most interesting or what depth of content they should include. Writing a list of QAs challenges authors with questions, such as *what questions would readers ask?* and *how detailed should the answers be?* These questions allow authors to gain an understanding of the interests and knowledge of readers. *"Because I work on ML applications for molecular structures of drugs, I talk to doctors and pharmacists a lot. Most of them don't have ML backgrounds. I wonder what they'd want to know about my research and how much I should go into detail about ML. (P2)"* Participants agreed that understanding readers' perspectives is critical to designing the *right* questions and curating their answers to enhance readers' understanding, including whether supplementary examples are needed or jargon is preferred.

Participants also raised concerns about specifying target readers and their interests in advance. They discussed the unspecified nature of readers online, such that social media services are typically open to anyone; thus, they shared the importance of taking multiple perspectives when crafting QAs to cover a broad range of readers. However, this strategy could be challenging. *"It will be great if people other than researchers read my Twitter posts on my research. But, I am not sure how much more information is needed for novices to make sense of my research. (P1)"* Moreover, there is a limited chance to learn about readers because of readers' reluctance to ask questions online asynchronously. Participants wish they could actively communicate with readers about their research work on social media, like Twitter, but they find people rarely leave comments on their research posts or send them direct messages. *"If you're sitting right next to the author during a conference, you can ask any question you like, even the most basic. But people don't ask questions on Twitter or any other online channel because they don't know if the authors are available. (P5)"*

3.2.2 DR2: Constructing Concise yet Informative QAs with a Narrative. Participants considered readability the most important factor for promotional content to attract readers' attention. They define highly readable promotional content as short writing that is easy to comprehend. Indeed, all participants mentioned an internet slang word, *Too Long; Didn't Read (TL;DR)*, emphasizing that most readers prefer a short summary to a long, detailed post. Participants reported that potentially, because of this negative sentiment against reading long texts, most promotional posts on social media these days tend to notify that a research article has been published, attaching a link to the article. However, merely making promotional content short is insufficient for pitching a research article; for example, *"effective promotional content at least describes how their research was done, what the findings were, and what arguments authors make. (P6)"* Other participants (P1, P3, P4, P7) also expressed frustration that most promotional posts on social media platforms (e.g., Twitter posts) and abstracts of research articles often fail to engage readers despite their short length as they are written presumably as part of common research practice without considering readers' interests deliberately. While participants agreed that presenting research articles using QAs would trigger readers' curiosity, they were concerned that simply listing QAs in a random order may not improve readability. They suggested that QAs should be presented cohesively as if they tell a well-written story.

Importantly, participants underscored the order of organizing QAs in providing a narrative of research work. As a research article is typically structured in a certain manner, such as starting from introduction, literature reviews, experiments, and results to discussion sections, formatting promotional content following the same order may not draw readers' attention. P8 mentioned, *"If QAs flow like a research paper, they may look boring. Authors need to craft QAs in a more interesting way. Perhaps showing the study highlights first can interest readers and lead them to continue reading QAs."* P3 suggested, *"You may start with easy questions and then move on to difficult ones for those who become more interested in your research."* While all participants were aware that ordering QAs is critical for engaging readers in their research work, they found scaffolding is necessary to assist authors in creating QAs in an engaging order. In fact, they mentioned the very first QA is important as it can serve as a hook to a story, but simultaneously, they struggled to develop one for their research article.

3.2.3 DR3: Speaking for Researcher Identities. Participants expressed a desire to grow in their research careers by promoting research articles. They hoped their promotional content would connect them with fellow researchers inside and outside their research domain and ignite discussions on the promoted research article (P1, P3, P5, P7). This way, authors can receive constructive feedback on their research, which will help them further develop their research (P1, P2, P3, P5). In addition, they would reveal tacit knowledge, such as behind-the-scene information or lessons learned from the study that did not appear in the paper. P3 mentioned, *"Researchers in my fieldwork with young patients, especially children with cancers. It is tricky to recruit study participants for various reasons like ethical and legal issues, not to mention persuading parents. Running studies with children itself is also challenging. Children have a relatively less attention span, and so on. (...) I believe sharing such information is helpful for fellow researchers in similar domains. Unfortunately, research articles cannot afford this information. It seems necessary to have an additional channel instead. (P3)"*

They further envisioned using the QA-style promotional content to solidify their identity as researchers. Most participants expected to build a presence within the research community by connecting with other researchers who have similar research interests (P1, P2, P5, P7). P1 mentioned, *"By promoting my research article, I would like to share my research interests and future plans."* The promotional content they suggested can hint to readers about the authors' research interest as well as their future research direction, which ideally leads to future opportunities, such as collaborative research projects and job positions. Their needs also expand to branding them as researchers. For example, P5 said, *"I am advertising my research findings. So, some researchers may approach me right after* *for potential collaboration [after reading my QAs], which is great. But in the long term, I hope the readers have a good impression of me that I am an expert on X and have successfully finished a project. And hopefully, they find me for collaborative research projects or job offers even after I leave my current program. (P5)"*

4 AQUA

Based on the design requirements, we developed AQUA, a co-creation system for researchers to write QA-style articles with LLMs to promote their research papers. AQUA was designed as a research probe for the purpose of understanding how users interact and perceive LLMs in co-creating QA-style articles [26].

We define key terms as follows. Firstly, **a QA pair** is a question and its corresponding answer. It is a basic building block of a story. We envision delivering a story by presenting QA pairs in a particular order. Secondly, **a QA-style article** refers to a one-pager that lists a series of QA pairs regarding a research paper. It is noteworthy that a QA-style article is concerned with only one research paper specified by its author(s). This one-pager can be shared as a URL. Lastly, **authors** are researchers who authored a published academic paper. They create QA-style articles for their research papers using AQUA. Authors may have more than one QA-style article. As authors are our target users, we use *users* and *authors* interchangeably in the remainder of our paper.

4.1 System Overview

To create a QA-style article, users select a research paper they published and start creating QA pairs in the editing space, called the QA Editing Space (Fig 1a). Users can either write questions by themselves or select one from recommendations. When users add a question, a corresponding answer is generated automatically. They can modify the automated answer. Users can create as many QA pairs as they want in the QA Editing Space, and they select QA pairs that will be displayed in the QA-style article. They can also change the order of QA pairs.

We utilize two LLMs to generate question recommendations and answers. For question recommendations, GPT-4 by OpenAI [47] was used. The details of prompt engineering are described in the *Features Leveraging LLMs* section. For answer recommendations, we used the Flan-T5-3B model [15] in the setup proposed in QASA [35], which demonstrates high accuracy in question answering in science domains. AQUA's interface is built using React, an open-source front-end JavaScript-based library. QA pairs in the QA Editing Space and those selected for the QA-style article were saved in the Google Firebase database [22]. All user interactions (e.g., button clicks for question recommendations) are logged onto the database.

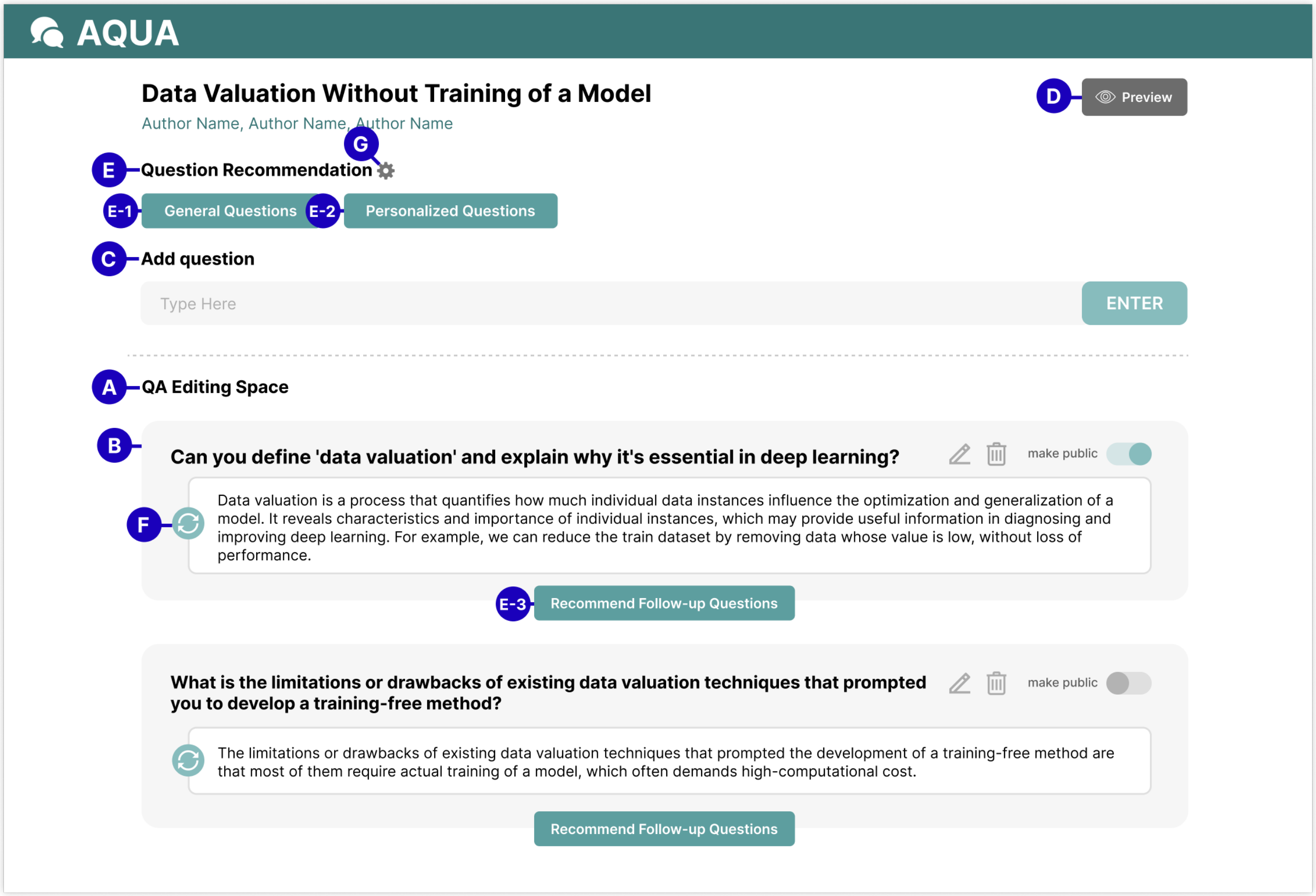
4.2 Key Features Leveraging LLMs

4.2.1 Question Recommendations. Users can request recommendations for potential questions. AQUA recommends three types of questions: (1) general questions, (2) personalized questions, and (3) follow-up questions. These three types of recommendations were designed to meet the design requirements (DRs) driven by the formative study. The details of each type, including its function, interaction design, and targeted DRs, are described below.

(1) **General question recommendation** suggests questions that probe into the content of a research paper, focusing on delivering essential pieces of information for readers to understand the research paper. General questions are asking about research motivation, research topic, methodology, results, and applications. **Interaction design:** As shown in (Fig. 2), users can get five general questions by clicking on the "General Questions" button. They can add questions by clicking on "Add+" displayed next to each question. Users can decide not to select any of the five and receive a new set of recommendations by clicking the "General Questions" button again. **DRs:** These questions simulate questions readers may ask in order to understand the paper; therefore, they provide users with a rethink about their own papers from the perspective of readers (DR1: Understanding perspectives of readers). By giving multiple questions at a time and letting users choose which one to add in any order they like, users are expected to experiment with developing a narrative that suits presenting their paper (DR2: Constructing concise yet informative QAs with a narrative).

(2) **Personalized question recommendation** suggests questions specific to users based on their research backgrounds, such as published research papers other than the one that is currently being promoted or the subject of the QA-style article. Unlike general questions that stick to the paper content itself, personalized questions ask about researcher-side stories, such as their motivation for research, future research plans, and trial-and-error during the research project. **Interaction design:** Users follow the same procedure as they would for general question recommendations except that the button is "Personalized Question" (Fig. 2). **DRs:** personalized questions lead to sharing something important that may not necessarily be included in the paper but can be heard only by the author, which we aim to address (DR3: Speaking for Research Identities)

(3) **Follow-up question recommendation** suggests questions based on an already created QA pair. Follow-up questions arise in several cases in natural settings. Readers realize they lack information to fully understand the answer in the QA pair, so they ask about the information. Readers may develop the discussion after reading the answer by asking for further information. **Interaction design:** Users receive three follow-up question candidates by clicking on the "Recommend Follow-up Questions" button, which is located on the bottom of a QA card (Fig. 3). They can regenerate them by clicking on the same button. They can select one question by clicking on the question, which will create a new QA card. **DRs:** The process of selecting follow-up questions contributes to developing a narrative (DR2: constructing concise yet informative QAs with narrative) by taking the perspectives of potential readers (DR1: understanding perspectives of readers). For example, only after reading a QA pair about study outcomes might users find that the assumptions of the study need to be explained to readers unfamiliar with the research topic so that the readers can understand the study outcomes.



Data Valuation Without Training of a Model

Author Name, Author Name, Author Name

Can you define 'data valuation' and explain why it's essential in deep learning?

Data valuation is a process that quantifies how much individual data instances influence the optimization and generalization of a model. It reveals characteristics and importance of individual instances, which may provide useful information in diagnosing and improving deep learning. For example, we can reduce the train dataset by removing data whose value is low, without loss of performance.

What is the limitations or drawbacks of existing data valuation techniques that prompted you to develop a training-free method?

The limitations or drawbacks of existing data valuation techniques that prompted the development of a training-free method are that most of them require actual training of a model, which often demands high-computational cost.

Figure 1: Main interface of AQUA provides "QA Editing Space (A)" for users to create and edit each "QA Card (B)," which is a visual representation of a QA pair. Users can edit questions and answers and set the visibility of QA Cards. Users can create a QA card by entering a question in "Question Input Box (C)". AQUA offers "Question Recommendation (E)," which comprises "General Questions (E-1)," "Follow-up Questions (E-2)," and "Personalized Questions (E-3)." Users can have their answers recommended by clicking the "Regenerate Answer (F)." It also allows users to prompt question recommendations features ("Additional Prompt (G)"). Users can click on the "Preview Button (D)" to overview their QA-style articles as shown to readers.

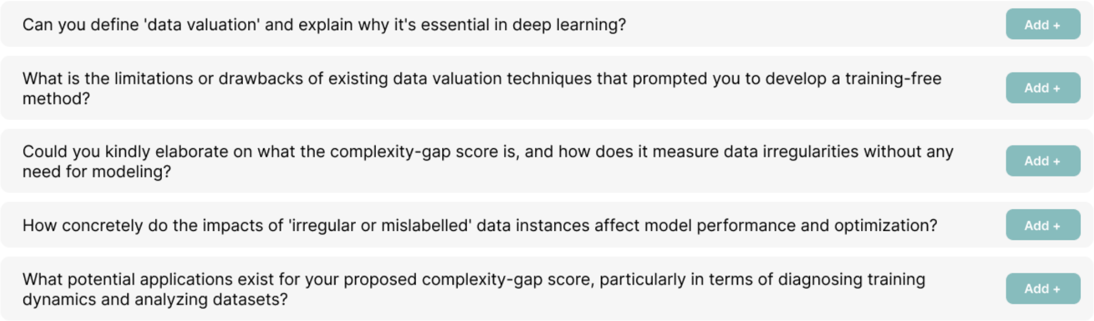
4.2.2 Customized Prompt Engineering. If users are unsatisfied with the recommended questions overall, they can provide additional prompts about what they would expect from question recommendations (e.g., tone, primary focus). Users open up "Additional Prompt," which pops up and asks for text input (Fig. 4). This feature allows users to conduct prompt engineering directly to guide underlying LLMs in AQUA. AQUA will recommend questions reflecting the updated LLMs. We intend to give more control to users in generating questions so that they can explore diverse narratives and build one they like (DR2: constructing concise yet informative QAs with narrative).

4.2.3 Answer Recommendations. AQUA drafts an answer to the question added by users. Once users add a question, an answer is generated and shown immediately. Answers are generated primarily on the research paper. Therefore, they may provide no information for questions whose answers cannot be found directly from the paper. Users can regenerate a new answer by clicking on the "Regenerate Answer" button at the bottom right corner of the QA card (Fig. 1f). Although the generated answers are not long, typically less than five sentences, users can make them shorter. Also, they can fix incorrect information, which comes from hallucinated text results generated by LLMs (DR2: constructing concise yet informative QAs with narrative).

4.3 Prompt Design for Question Recommendations Features

In this section, we describe how we design prompts, which are text inputs to guide LLMs to generate expected outputs, for our question recommendation features.

As illustrated in Fig. 5, all three question types share the same fundamental prompt to generate questions in a coherent context and tone. In the fundamental prompt, we specified a persona and a



Can you define 'data valuation' and explain why it's essential in deep learning?

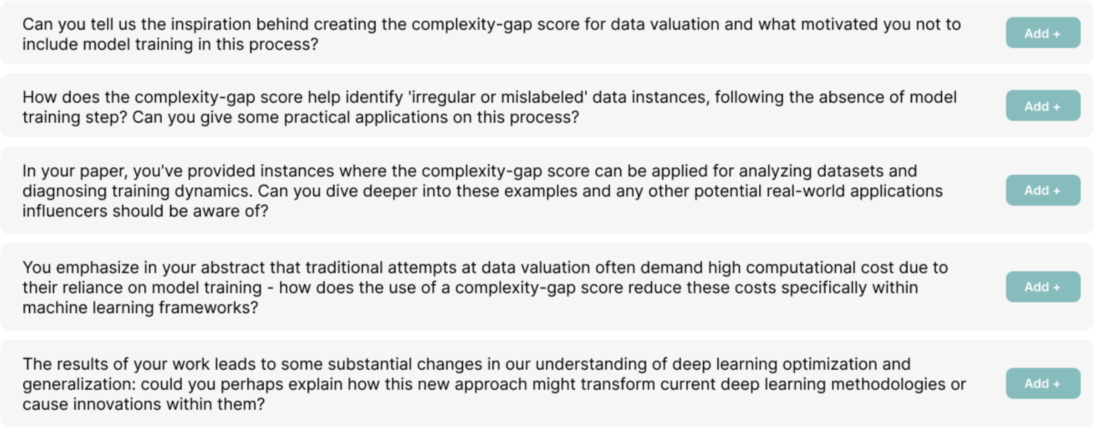
What is the limitations or drawbacks of existing data valuation techniques that prompted you to develop a training-free method?

Could you kindly elaborate on what the complexity-gap score is, and how does it measure data irregularities without any need for modeling?

How concretely do the impacts of 'irregular or mislabelled' data instances affect model performance and optimization?

What potential applications exist for your proposed complexity-gap score, particularly in terms of diagnosing training dynamics and analyzing datasets?

(a) Examples of General Question Recommendations from P12's experiment



Can you tell us the inspiration behind creating the complexity-gap score for data valuation and what motivated you not to include model training in this process?

How does the complexity-gap score help identify 'irregular or mislabeled' data instances, following the absence of model training step? Can you give some practical applications on this process?

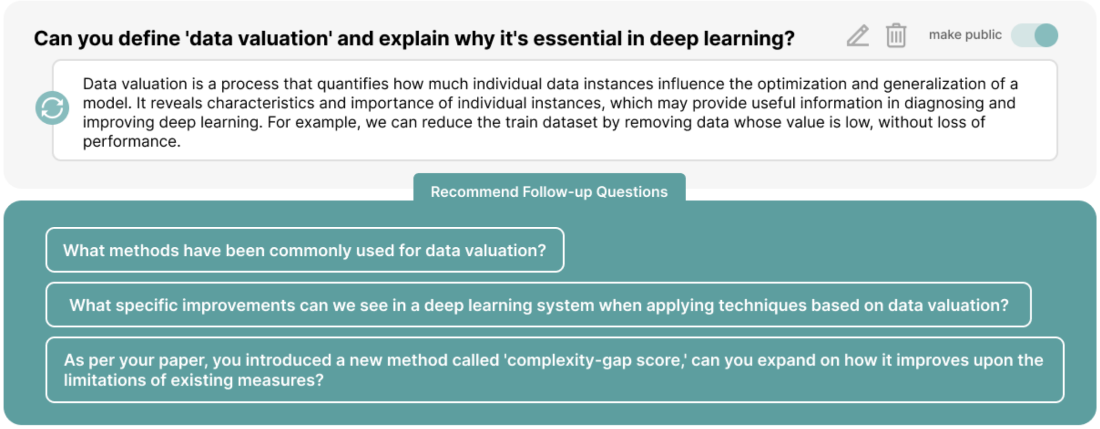
In your paper, you've provided instances where the complexity-gap score can be applied for analyzing datasets and diagnosing training dynamics. Can you dive deeper into these examples and any other potential real-world applications influencers should be aware of?

You emphasize in your abstract that traditional attempts at data valuation often demand high computational cost due to their reliance on model training - how does the use of a complexity-gap score reduce these costs specifically within machine learning frameworks?

The results of your work leads to some substantial changes in our understanding of deep learning optimization and generalization: could you perhaps explain how this new approach might transform current deep learning methodologies or cause innovations within them?

(b) Examples of Personalized Question Recommendations from P12's experiment

Figure 2: AQUA presents question recommendation features at the top of the user interface. Depending on the button the user clicks, the system provides five questions that are either general (a) or personalized (b). Both examples were excerpted from P12's experiment.



Can you define 'data valuation' and explain why it's essential in deep learning?

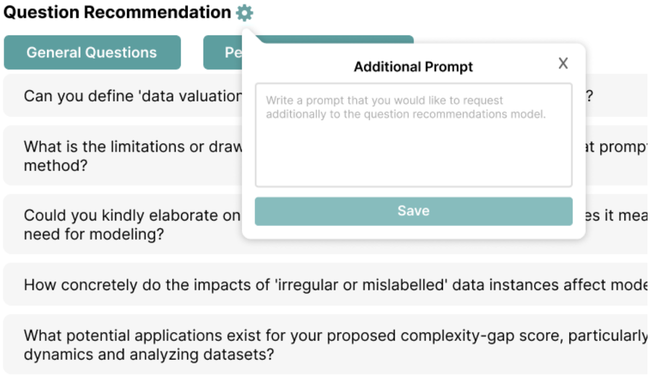
Data valuation is a process that quantifies how much individual data instances influence the optimization and generalization of a model. It reveals characteristics and importance of individual instances, which may provide useful information in diagnosing and improving deep learning. For example, we can reduce the train dataset by removing data whose value is low, without loss of performance.

What methods have been commonly used for data valuation?

What specific improvements can we see in a deep learning system when applying techniques based on data valuation?

As per your paper, you introduced a new method called 'complexity-gap score,' can you expand on how it improves upon the limitations of existing measures?

Figure 3: AQUA suggests potential follow-up questions for each QA pair. When users click on the "Recommend Follow-up Questions" button located at the bottom of each QA card, the system displays three follow-up questions under the QA card.



Can you define 'data valuatior ?

What is the limitations or draw it promp method?

Could you kindly elaborate on es it mea need for modeling?

How concretely do the impacts of 'irregular or mislabelled' data instances affect mode

What potential applications exist for your proposed complexity-gap score, particularly dynamics and analyzing datasets?

Write a prompt that you would like to request additionally to the question recommendations model.

Figure 4: AQUA provides a collapsible UI where users can add a prompt to question recommendations.

research paper; we set up a podcast host who interviews research paper authors as the persona. Also, we provided the title and abstract of the research paper. We chose not to provide the entire paper since we observed it yielded questions that were too complex for first-time readers.

In addition to the shared fundamental prompt, we added additional prompts catering to the types of questions. The details of prompt designs are as follows.

4.3.1 Prompting general questions. We referenced interview-based research podcasts where a host interviewed scholars on their research. Two authors of this paper randomly selected 20 episodes of Microsoft research podcasts [53] released over the past two years and analyzed the scripts together to identify themes that frequently appeared when scholars explained their research papers. The five most frequent themes were motivation, background, experimental details, results, and future work. We added descriptions of these five themes to guide LLMs in creating questions covering them. If users have additional requirements (see *Customized Prompt Engineering*), they are added to the end of the current prompt.

4.3.2 Prompting personalized questions. Additional information other than the selected research paper is necessary to include in order to enrich personalized questions. We added users' research backgrounds, such as research agendas and topics. We used Semantic Scholar [56], a web application that manages and summarizes scholarly work, to retrieve users' past publications. The authors' research background information was made out of past publications and added to the prompt.[1] Similar to prompting general questions, users' additional requirements are added to the end of the prompt if there are any.

4.3.3 Prompting follow-up questions. We included the QA pair in the prompt to allow LLMs to recommend follow-up questions for that pair. Then, we added a request in the prompt to recommend three follow-up questions that could follow that QA pair.

Participant ID	Age	Gender	Current Status	Field
P1	28	F	Ph.D	AI/ML
P2	23	M	Undergraduate	AI/ML
P3	29	F	Ph.D	AI/ML
P4	20	M	Undergraduate	AI/ML
P5	27	F	Master	HCI
P6	27	M	Ph.D	AI/ML
P7	25	F	Master	AI/ML
P8	30	M	Ph.D	AI/ML
P9	28	M	Ph.D	AI/ML
P10	24	M	Ph.D	AI/ML
P11	28	M	Ph.D	AI/ML
P12	24	M	Ph.D	AI/ML

Table 1: Demographic information of study participants.

5 USER STUDY

Our study was designed to explore how authors interact with and perceive the LLM-powered features of AQUA in creating QA-style articles. Instead of conducting a comparative study to evaluate each of the system components, we conducted a single user study to closely observe authors' interactions with AQUA, thereby gaining a more in-depth understanding of the interactions between humans and LLMs. This study was approved by the university's institutional review board (IRB).

5.1 Participant Recruitment

A total of 12 participants were recruited from research communities associated with universities in South Korea, excluding those who took part in the formative study. All participants were early-career researchers who had papers published in AI/ML and HCI domains as the first author and had experience promoting their papers through social media platforms. Their demographic details are presented in Table 1. Each participant received $40 as compensation for their participation.

5.2 Procedure

Before the study, participants received pre-study guidelines, which included instructions to create a Semantic Scholar account and select one of their published papers on which they would write a QA-style article. At the beginning of the study, we explained the study's purpose to the participants and provided a detailed walkthrough of AQUA's features. As the main task of the study, participants created a QA-style article following three instructions: (1) Create a polished QA-style article that they could share online to promote their work; (2) Choose the number of QA pairs that can best present their papers in a sensible length; (3) Take as much time as they need. After completing the main task, participants self-evaluated their QA-style articles on a 7-point scale regarding readability, conciseness, clarity, attractiveness, coverage, depth, and uniqueness, which are commonly assessed qualities of writing [11]. Lastly, participants had a 40-minute semi-structured interview during which they elaborated on their QA-style article and shared their experience of using AQUA.

---

[1] In our user study, we asked participants to share their Semantic Scholar author page URL before the study. Then, we used ChatGPT to write a short research background by referring to the publication list on Semantic Scholar.

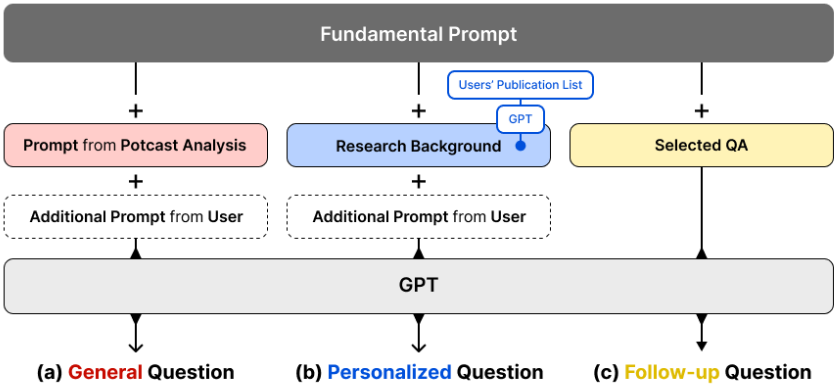


Figure 5: Each question creation function operates based on the fundamental prompt. General Question (a) is a prompt created based on podcast script analysis and recommends general questions needed for users to introduce the paper. Personalized Question (b) recommends a personalized question to the researcher based on the researcher's background. Follow-up Question (c) recommends a question that may follow the QA pair.

5.3 Analysis

We analyzed four types of data: QA-style articles, authors' self-evaluation of articles, log data, and interview data. Firstly, we conducted a descriptive statistical analysis of the QA-style articles created by participants. To further understand the content of articles, we classified QA pairs based on an existing question taxonomy proposed for academic writing context, which consists of categories including opinion, result, system, application, method, and aim [36]. We only classified the QA pairs that authors chose to include in the final version of the articles. Two authors independently classified and then reached a consensus on the final classification through iterative discussions. Additionally, we analyzed authors' self-evaluation of writing qualities to understand how they perceived the QA-style articles they created using AQUA.

Next, we analyzed users' log data for the LLM-powered features. For question recommendation features, we obtained (1) the frequency of clicks on each of the three QA recommendation buttons, (2) the number of recommended questions accepted by users, and (3) the number of recommended questions that were put on the final QA article. For the answer recommendation feature, we investigated the number of QA pairs whose answers were revised. Then, we categorized the types of revisions made by participants. We also obtained the number of times additional prompts were provided by users.

Lastly, we qualitatively analyzed the interview data using thematic analysis [10]. The first author initially open-coded the interview transcripts and observational notes using Altas.ti [3], iterating several times. The entire research team then discussed and identified patterns and themes through multiple rounds of group meetings.

6 FINDINGS

6.1 Descriptive Summary of AQUA Usage

6.1.1 QA-style Articles. Our 12 participants created 13 QA-style articles for one of their papers through AQUA.[2] The average time spent writing an article was 40.3 minutes (SD = 11.50, Max = 57.8 minutes, Min = 35.0 minutes). A QA-style article contained an average of 6.85 QA pairs (SD = 7.16) ranging from 3 to 15 pairs (Fig. 6c). The average length of the questions was 14.97 words (SD = 6.40), and the average length of the answers was 54.72 words (SD = 22.58). Examples of QA-style articles written by participants are provided in Appendix A.

The classification of the questions sheds light on authors' strategies for presenting research articles by showing what type of content they select for promoting their research (See Table 2). The most frequently appearing category of questions was "System" (22.47%), which mainly included questions related to the principles of the system and distinctions from existing systems. The least common category was "Opinion" (11.24%), which mainly included questions about the authors' subjective opinions or claims. Some questions did not fit well in the taxonomy, and they were mostly written about summaries of papers or background knowledge.

We also wanted to comprehend how participants evaluate their QA-style articles, so we asked them to rate their articles according to seven writing qualities (Fig. 7). The highest ratings were for readability (M = 5.92, SD = 0.51) and clarity (M = 5.50, SD =1.00), while the lowest were for uniqueness (M = 2.75, SD = 1.36) and depth (M = 3.83, SD = 1.64).

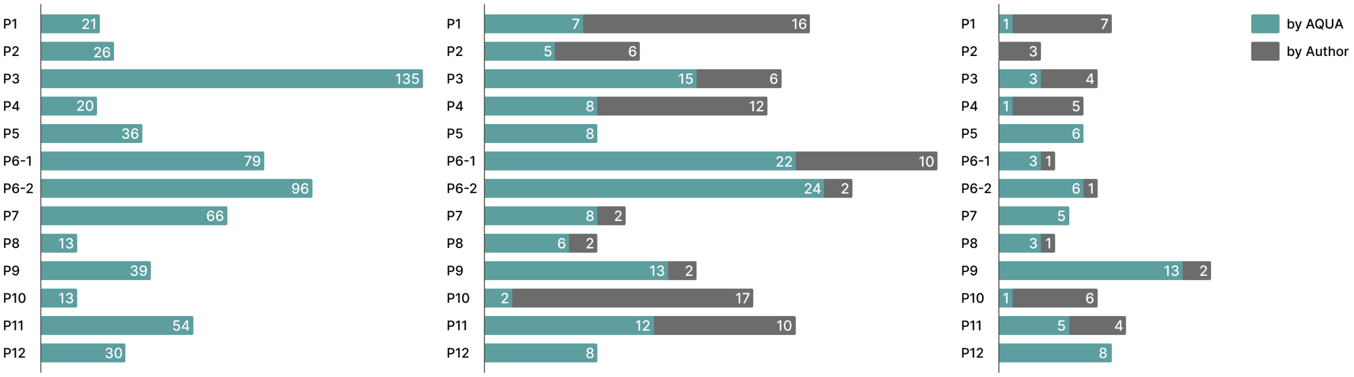
6.1.2 The Usage of Key Features.

Question Recommendations. On average, participants used question recommendation features 11.5 times (SD = 10.40), and they received 48.30 recommended questions (SD = 42.20) (Fig. 6a). Approximately 22% of the recommended questions (M = 10.60, SD = 8.41 questions) were added to the QA Editing Space (Fig. 6b). Finally, approximately 40% of the added questions (M = 4.23, SD = 4.88 questions) were selected for the final QA-style articles (Fig. 6c).

The breakdown of usage by question type is presented in Table 3. Participants most frequently used the general question recommendation. They added 6.31 general questions (SD = 2.78) to the QA Editing Space and included 3.00 of these questions (SD = 2.94) in their QA-style articles. Conversely, participants added only 1.46

---

[2] P6 created QA-style articles for two papers, resulting in a total of 13 articles. This was driven by his curiosity about how AQUA would handle questions and answers for a simpler paper, as the first paper he chose was complex with many experiments.



(a) Number of Questions Recommended by AQUA (b) Number of Questions Added to Editing Space (c) Number of Questions Included in QA-style Articles

Figure 6: The numbers of recommended questions (a), added questions (b), and finally selected questions (c) for each participant. The usage of AQUA's question recommendation features varied by participant.

Category	Count	Example Questions
Opinion	10	What is the implication of this paper for GNN practitioners? *(Generated by Author, P6)* Could you discuss the existing shortcomings in traditional traffic prediction models that led to your research? *(Generated by General Question, P9)*
Result	12	How much has AltUB improved the performance of anomaly detection models? *(Generated by Author, P4)* Your framework was tested on MNIST and CelebA datasets, what were these experimental findings or observations? *(Generated by General Question, P7)*
System	20	How does MED-SE work? *(Generated by Author, P2)* How does SuperGAT's ability to distinguish improves representation particularly when faced with a noisy graph? *(Generated by Follow-up Question, P6)*
Application	13	What could be the potential applications of this research for both NLP and Political fields? *(Generated by Author, P3)* What are potential future research directions opened by introducing human activities into traffic pattern analysis? *(Generated by General Question, P9)*
Method	12	Can you describe the method proposed by the authors? *(Generated by Author, P10)* Can you explain the experimental setup that TAST tested? *(Generated by Author, P11)*
Aim	13	What is the main contribution of AltUB? *(Generated by Author, P4)* Could you explain further about how the 'Complexity-Gap Score' proposes its unique implications for finding 'irregular or mislabeled' data instances? *(Generated by Personalized Question, P12)*
Other	9	Summarize this paper in 50 words! *(Generated by Author, P1)* Could you provide some background on normalizing flow and its significance in unsupervised anomaly detection? *(Generated by Author, P4)*

Table 2: The questions in the QA-style articles created by the user are categorized according to a taxonomy of questions.

personalized questions (SD = 2.22) to the QA Editing Space and accepted 0.38 of these questions (SD = 0.87). Only three participants (P5, P6, P12) included the personalized questions in QA-style articles. Additionally, participants added 2.84 follow-up questions (SD = 3.41) to the QA Editing Space, and accepted 0.85 of these questions (SD = 1.07), with only five participants(P3, P5, P6, P9, and P11) accepted the follow-up questions.

Participants added 6.54 questions (SD = 5.92) by themselves. Notably, P1 and P10 each added more than 16 questions directly, while P5 and P12 did not add any questions themselves (See Fig. 6(b)). Participants included 2.61 of these questions (SD = 2.28) in QA-style articles.

Use of Additional Prompt. Half of the participants used "Additional Prompts." Six participants (P1, P3-6, P9) provided additional guides, totaling 11 instances. There are two types of inputs: (1) requests specifying the scope of the question (7 instances) and (2) requests specifying the tone of the question (4 instances). Prompts that specify the scope of the question include *"Please give me a question that relates to the application of this paper (P3)."* and *"Focus on preliminary experiments that the authors did to find the optimal attention design (P6)."* Examples of prompts that specify the tone of the question were *"Please answer within three sentences (P3)."* and *"Create a question that an elementary school student would ask (P5)."*

Answer Generation. On average, each participant edited 3.46 QA pairs (SD = 3.15) among the QA pairs in their QA-style articles.

	Button clicks	Recommended	Questions Added	Accepted
General question	M=4.69 (SD=3.09)	M=23.50 (SD=15.50)	M=6.31 (SD=2.78)	M=3.00 (SD=2.94)
Personalized question	M=2.15 (SD=2.44)	M=10.80 (SD=12.20)	M=1.46 (SD=2.22)	M=0.38 (SD=0.87)
Follow-up question	M=4.69 (SD=4.85)	M=14.10 (SD=14.56)	M=2.84 (SD=3.41)	M=0.85 (SD=1.07)
User-written question	-	-	M=6.54 (SD=5.92)	M=2.61 (SD=2.28)

Table 3: "Button clicks" refer to the number of times users clicked on question recommendation buttons. "Recommended" refers to the number of questions shown to users as the result of clicking on a question recommendation button. "Added" refers to the number of questions users chose and added to the QA Editing Space. "Accepted" refers to the number of questions users finally accepted to their QA-style article.

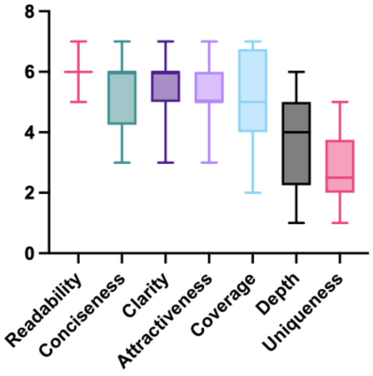


Figure 7: Results of participants' self-evaluations of the QA-style articles they created through AQUA.

They revised the recommended answers from about half of the QA pairs listed in the QA-style article (M = 51.27%, SD = 35.98%). While P2 revised all the QAs (3 out of 3) and P12 revised 87.5% (7 out of 8) of the QAs in their articles, P1 and P10 did not revise any QAs.

We analyzed how participants modified recommended answers and deduced revision patterns in context. (See Table 4). The predominant pattern was correcting errors in LLM-recommended answers. The second most frequent pattern was rewriting answers to enhance clarity and readability, although the recommended answers were not incorrect. Participants added details or clarified ambiguous parts to make their answers as clear as possible. Moreover, participants changed terms or phrases to reflect their intended tone and mood beyond focusing on the answer's content. Lastly, participants wrote the entire answer to questions about authors' opinions or future applications, which were out of the scope of the research paper. They had to do so partly because AQUA was designed to generate recommended answers based on the research paper, and therefore, it provided abstract answers or even no answers to such questions.

6.2 How do authors interact with LLM-powered features when creating QA-style articles?

6.2.1 Use Patterns of AQUA for Question Generation.

Waiting for Gotcha Moments for Better Questions. We observed that our participants repeatedly clicked on the question recommendation buttons. Participants (P3, P4, P5, P6, P7, P9, P11, and P12) expected the LLMs would keep drawing relevant questions at random and requested recommendations repeatedly until they received valuable questions. They did so either to select a recommended question when they found the *best* one or to write a question on their own inspired by recommended questions. These participants were impressed when LLMs generated insightful or critical questions. *'I was shocked when LLMs recommended a question about the computational cost. My research involves a technique of attaching multiple modules, which typically implies longer learning times. So, it(recommend question) is a common question I face while presenting my research. While I'm not sure if the LLMs deduced this, I was impressed that it brought this up as a follow-up question. (P11)"*

Furthermore, participants received recommendations to write more engaging and readable questions by themselves looking for expressions. *"Originally, I clicked on 'General Question' multiple times to receive recommendations for different types of questions. The one I got was about the same motivation as the one I had already added, but with much better language. So, I incorporated that phrasing and kept clicking for more recommendations until better expressions emerged, too. (P7)"* Other participants were just looking for ideas. *"To be honest, I rarely used the recommended questions directly, but they [generated questions] gave me ideas for other questions. (P4)"*. P11 also said, *"I always clicked on the follow-up question button after submitting a QA pair. Though I didn't implement them, I used them to inspire other related questions. (P11)"*

Developing a Narrative through Asking Follow-up Questions. Since AQUA automatically recommends answers, some participants realized they could exchange questions and answers as if they were having a conversation with AQUA. Some participants (P1, P4, and P10) were deeply engaged in making QA pairs as if they had a conversation with the system. For example, P1 came up with follow-up questions based on the LLM's responses, naturally engaging in a question-and-answer exchange with the LLMs.

```
Question by P1: Explain the key methods of
this paper in 50 words!
Generated answer: In this paper, the authors
used machine learning models and data
augmentation to predict well-validated fMRI
markers of human cognition from multivariate
patterns of fNIRS. (. . . )
Question by P1: I see - what are the fNIRS
and fMRI then? What are the similarities and
differences?
```

Category	Recommended Answer from LLMs	Edited Answer
Error	TAST-BN is a variant of TAST that fine-tunes the nearest neighbor (BN) layers instead of projection heads. (...)	TAST-BN is a variant of TAST that fine-tunes the batch normalization (BN) layers instead of adaptation modules. (...)
Detail	(...) Additionally, all three models produced similar IS and FID scores, indicating that the framework can successfully unlearn the target feature while maintaining high-quality image generation.	(...) Additionally, all three models (Original model, unlearned model, baseline model) produced similar IS and FID scores, indicating that the framework can successfully unlearn the target feature while maintaining high-quality image generation.
Clarity	(...) because the classical approach often yields unstable results.	(...) because the classical approach often fails to convert images into the base distribution.
Tone	The authors developed a new framework (...)	We develop a new framework (...)
Beyond Paper	The paper does not provide specific information on possible extensions of the complexity gap score's (...)	The feature complexity gap score seems to be a very effective score, and can be applied in the real world scenario. (...)

Table 4: Examples of Participants' LLM-Generated Answers Revising Patterns.

P1 continued to add questions as if asking questions directly to the system, probing into the parts of the paper where the generated answers were lacking until the QA pairs covered the entire paper comprehensively. *"First, I asked for an explanation of the method. But it didn't say which model was used, so I asked again. This time, it didn't mention that we augmented the model through brain imaging, so I inquired about that. The reply didn't mention the task of taking brain images, so I then asked about that. (P1)"* After the conversation, they developed their own narratives by revising a conversation-like list of QA pairs.

Creating and Refining Questions by Themselves. A few participants (P2, P8) made the QA pairs by themselves without using question recommendation features. For example, P2 had already decided what questions the article should consist of in order to present the paper and barely used the system's question recommendation feature. *"The promotional article for the research paper I saw on Twitter was usually divided into eight or nine slices, but I thought there were too many. I organized the questions like this because, from the perspective of other researchers, I thought they would be most curious about 'why we need this', 'how this works,' or 'what the results are.' (P2)"* P8 only used the recommended questions that matched his initial narratives. He mentioned that he would have used the question suggestion feature more if it had offered him more unexpected and creative questions.

6.2.2 Usage Patterns of AQUA for Editing Generated Answers.

Building Trust. Most participants (11/12) emphasized the importance of providing accurate answers in research promotional content. They were aware that answers recommended by LLMs may contain incorrect information due to a widely known phenomenon of LLMs' hallucinations–producing factually wrong or misleading information not supported by authors' paper [75], so they considered verifying generated answers of vital importance before publishing them to audiences. Some of them even added several questions completely irrelevant from the research context just to test how trustful LLM-generated answers were.

```
Question by P4: Which menu do the authors
recommend for the lunch?
Answer: The paper does not provide
information on a specific menu for lunch.
```

Through this process, participants built trust in the answers recommended by AQUA. They tend to accept the recommended answers when their trust in the system's recommendation capabilities is reinforced through these verification processes. However, the system quickly lost their trust once any answer contained incorrect quantitative data or key terms presented in the paper.

Barely Reviewing Answers. Although AQUA allows and encourages participants to edit answers, many participants (7/12) did not edit the answers at all or edited them minimally. Some participants might not thoroughly review and edit the generated questions potentially because of their strategy of creating QA-style articles such that they added a number of QA pairs first without carefully reviewing answers and then selected QA pairs out of the long list based sorely on questions. We suspect that these participants felt too overwhelmed to review individual answers after creating a number of QA pairs. P3 said, *"After creating all the QA pairs, I found the number of answers overwhelming to revise in detail, so I ended up submitting most of them in their original form."*

Since revising answers takes significant effort, participants tended to avoid revising generated answers if they were factually correct. For example, P1 did not revise the recommended answers if they were not incorrect, although she was not satisfied with their personality and nuance because it would take a long time to revise them. Some participants (2/12) improved questions to let AQUA generate modified answers for them rather than revising answers by themselves. P10 mentioned, *"If I didn't like an answer, instead of revising it, I asked the question again until satisfied."*

6.3 How do authors perceive LLM-powered features when creating QA-style articles?

6.3.1 Reducing Burden on Authors. Participants (9/12) reported that LLM-powered features alleviated the challenge of the cold start. These features served as guidelines for those who were unsure about how to start writing QA pairs. Most participants (11/12) began by clicking on the "General Question" button to receive a list of recommended questions, and they started creating their QA pairs by selecting questions from the list. Moreover, most participants (10/12) appreciated the answer generation feature for drafting answers automatically for them, although they perceived the generated answers as imperfect. *"Although the answers were not correct, it gave me an outline of how I should write the answers, which was quite useful for me. (P12)" "Without Answer Suggestion, I would have had to check my paper to give a correct answer, but as this gave me the sources, it was much easier to write an answer. (P10)"*

6.3.2 Helping Authors Take Readers' Perspectives. Overall, participants shared that AQUA helped them create a QA-style article from the perspective of readers. For those who are already familiar with the paper's content, it is easy to overlook the viewpoints of first-time readers who may not have prior knowledge sufficiently. After looking through questions recommended by AQUA, participants (6/12) realized that explaining about the fundamentals of their studies, such as study design, is essential for delivering the gist of the paper. For example, P6 mentioned, *"When I started brainstorming QAs for my paper, I focused on delivering the findings and implications of my study. But soon after I had question recommendations, I realized I also needed to clarify how my experiment was designed and carried out. The task I did was recently introduced in the field like three years ago, so people would be curious about why I chose the task instead of classical ones. I was impressed AQUA picked this up."* Moreover, participants (3/12) were also surprised that the recommended follow-up questions matched the ones they were asked by fellow researchers. *"It's surprising that the system recommended the same question I was asked at a conference presentation. This research is still in progress, so the audience at the conference was curious about future work, and the LLMs asked about it as well. (P5)"*

6.3.3 Generating Questions Misaligned with Authors' Intentions. Participants were hesitant to add follow-up and personalized questions to their final QA-style articles for several reasons. In the case of follow-up questions, participants (6/12) thought that the questions were plausible questions following after the initial QA, but they were often too specific to put on the promotional page. P9 mentioned, *"The questions are really interesting and seem like the kind that could come up in a real-world conference. However, I believe that promotional content should only contain the key points of a paper, and these questions seem too minor to include in promotional content."* This specific nature of the follow-up questions misled participants, made them think the questions are irrelevant and not fulfilling their intention of creating a logical narrative. For instance, P5 mentioned, *"The recommended follow-up question differed from what I had expected. For example, I thought a follow-up question about the method would start with a 'why,' asking about the rationale of using a qualitative method. Instead, the system suggested questions somewhat irrelevant to the initial question, such as asking about the participants' demographics."* In the case of personalized questions, participants found it vague and tangential to the research paper. (e.g., asking about the motivation of pursuing a research career in X domain). Since personalized questions were generated based on their previous research, most of our participants who had limited research experience received vague questions and felt they did not receive meaningful questions. They mentioned they wish they were asked about their future research plans or goals in relation to the promoting research paper rather than their past work.

6.3.4 Useful in the Research Process Beyond Research Promotion. Participants (3/12) suggested that AQUA can further support various aspects of academic communications, such as improving manuscripts and practicing conference presentations. They recognized the potential of using the system to aid in refining the content of their papers in the pre-publication phase. For example, P3 felt that receiving questions from the system was similar to undergoing an open review process, therefore, using the system provides an opportunity for reviewing a written manuscript from a second eye and receiving feedback for improvement. P9 also highlighted that reviewing the LLM-recommended answers offered a chance to contemplate comprehensible ways to communicate their research with readers. Furthermore, participants mentioned that this system could be used to prepare conference presentations. P7 said, *"As I still have little experience at academic conferences, I am worried about what questions I would get and how I would answer when I give a presentation at a conference. The system's question recommendation feature is like experiencing a Q&A session at a conference in advance. It allows me to practice responding to anticipated questions."*

7 DISCUSSION

The present study sought to understand authors' use and perception of LLMs in writing for research promotion. To this end, employing AQUA as a research probe, we analyzed how authors interacted with LLM-powered recommendation features during the process of creating QA-style articles. We found that authors actively used the features and perceived them as helpful for initiating writing and gaining insights into readers' perspectives. These findings align with the positive roles of LLMs in the co-writing process reported in previous research [e.g., 23, 39, 77].

The analysis of authors' behavioral engagement with AQUA shows that authors perhaps overly rely on LLMs during the co-creating process. Authors expected LLMs to recommend the *right* questions to build a unique and insightful narrative. Most of them repeatedly requested sets of question recommendations instead of writing questions by themselves, even after learning that LLMs had limited reasoning capabilities sufficient enough to understand research papers and authors' preferred narratives. Additionally, authors minimally edited LLM-generated answers. While reducing their workload, this inadvertently left authors passive in the creation process. Contradicting our formative study findings that authors desire to embed their personal viewpoints and experiences when promoting their research, the QA-style articles created by our participants often fell short in revealing the authors' distinctive voices and deep insights.

We discuss how to design human-LLM interactions to foster authors' active participation in co-writing with LLMs and extend the use of LLMs in real-world research communication beyond creating research promotion content.

7.1 Beyond Simple Recommendations: Promoting Interactive Exchanges with Large Language Models

Our findings indicate that authors seek writing support from LLMs closely aligned with their intended narrative and writing style instead of merely suggesting a number of recommendations. In our context, individual authors had varied demands on LLMs when writing QA-style articles. Therefore, it is crucial that authors consistently convey their intentions to LLMs to have more appropriately tailored questions [41]. One notable approach is to allow authors to modify the LLMs' prompts directly. However, our findings revealed that asking authors to customize their own prompts can add an extra burden and interfere with their original work. An intuitive interface may be a valuable consideration to ease the authors' burden [72], as prior research suggested that the effectiveness of the interface in delivering users' intentions was key to successful co-drawing activities with LLMs. [16].

Furthermore, our findings suggest that interactions with LLMs extend beyond simple question recommendations. In this study, participants interacted with AQUA as if they were conversing with it to collaboratively create a QA-style article narrative, even though AQUA was not designed as a conversational agent. Notably, they found it beneficial to generate follow-up questions in a multi-turn dialogue with AQUA, allowing them to delve into the audience's perspective. Previous research supports this by showing that conversations with LLMs stimulate users' participation and interest, leading to creating more content [46]. Given that LLMs may not fully capture the author's diverse intentions in recommending questions, future systems could benefit from incorporating mechanisms that enable authors to generate questions through interactive dialogues with LLMs tailored to their needs.

The capability of LLMs to simulate specific personas can provide a novel approach for authors to craft QA-style articles [49], taking into account a variety of perspectives and needs, including those of their target audiences in research communication context [4]. Authors can get diverse questions by prompting LLMs to role-play as various stakeholders, such as the general public, a researcher from a different field, or a subject matter expert. Moreover, as our participants recognized LLM's answer recommendation feature as an alternative to themselves, LLMs can be used to simulate the authors themselves. By simulating a scenario where the authors are having conversations with their 'alternate self' who introduces the paper, they can think and ask questions about their papers from a third-party perspective. These simulation-based interactions can be significant in creating QA-style articles that originate from the conversation and can be even more emphasized in our context, where authors need to understand the perspectives of diverse readers and share their thoughts about the research.

7.2 Enhancing Author Engagement: Strategies for Ensuring Content Accuracy and Promoting Active Participation in LLM Interaction

Our findings indicate that although AQUA's answer recommendation features significantly reduce the effort authors need to put into writing answers from scratch, they still require authors to verify whether LLM-generated answers contain accurate information and align them with the intended tone and mood. Despite recognizing this need for revision, participants tended to trust the LLMs' ability to generate appropriate responses, often neglecting thorough reviews. Given the critical importance of conveying accurate information in promotional articles of research papers, it is essential to engage authors in meticulous review processes for content co-creation with LLMs.

Previous research indicated that user engagement in human-AI interaction can vary depending on the user's degree of trust and reliance on AI [31, 70, 76]. For instance, in tasks in video annotation, zealous AI, which is not confident in its response, has been observed to enhance human-AI team performance by increasing user involvement compared to restrained AI [70]. This is consistent with our finding that authors barely revised their answers when they built trust in the LLMs' recommendations. It indicates the intricate trade-off between an author's trust in the LLMs and their active participation, suggesting that effective co-creation results from carefully balancing them. Consequently, future research could investigate ways to improve user participation in interactions with LLMs by managing an appropriate reliance on LLMs that encourages user engagement.

Considering the substantial task of thoroughly reviewing all responses generated by the LLMs, we propose that the review process be streamlined by adjusting the level of trust depending on the type of response. Our study indicated that authors require only minimal or no revisions to LLM-recommended answers, which directly point to the content explicitly stated in their papers. On the other hand, answers related to subjective content, such as future work or the authors' subjective opinion not explicitly addressed in the papers, required their significant involvement. Therefore, LLMs can inform the author of a high confidence level for the former case, facilitating a streamlined review process. For the latter case, LLMs should signal a lower confidence level, prompting a more thorough review by the author. This approach aims to balance the efficiency of automated response generation with the necessity for accuracy and authors' voices in more subjective responses.

7.3 From LLM Interactions to Real-World Applications: Leveraging Large Language Models to Foster Academic Communication Skills

LLM-powered systems, such as AQUA, can effectively support researchers in enhancing their research communication skills. We found that our participants with limited experience in research presentations often feel apprehensive about Q&A sessions in academic conferences and recognize the need for practice. Through interaction with AQUA, these individuals had the opportunity to consider how to present their research in their own words and to prepare for potential questions from readers. Moreover, AQUA assists researchers in exploring different narratives, helping them extract and articulate their work's core ideas (e.g., through QA pairs) for promotional content. There have been ongoing efforts to utilize AI, including LLMs, for training required skills and facilitating academic communication, not only in written communication [39], such as writing promotional content but also in oral communication, such as giving presentations [21]. Future research could explore the use of LLMs in simulating different types of audiences. This would enable researchers to adopt varied perspectives and refine their strategies for promoting research. Additionally, such studies could examine how these simulated interactions enhance researchers' skills in engaging with diverse audiences, handling challenging questions, and presenting complex ideas with clarity during live presentations. As LLMs diversify their roles in supporting varied research tasks, we further envision expanding their capabilities in advancing research promotion and academic communication.

8 LIMITATIONS

While our findings provide insights into the interaction between authors and LLMs in generating QA-style research articles, this study has notable limitations. Firstly, our sample size was limited, and to generalize our findings more robustly, we intend to conduct experiments with a larger and more diverse participant pool in subsequent studies. Also, since we focused on early-stage researchers, further research is needed on the practices of senior researchers. Additionally, the preferences for promotional articles might vary based on the research field, and our study primarily targeted researchers in the AI/ML/HCI domains, potentially omitting perspectives from other disciplines. Lastly, our research focused on generating QA-style articles tailored to authors' desires. Future research should consider quantitatively evaluating the efficacy of these articles in aiding actual readers.

9 CONCLUSION

In this study, we propose AQUA, a research probe that enables researchers to co-create QA-style articles with LLMs to promote research papers. Using AQUA as a probe, we examined how users interact and perceive LLMs in co-creating QA-style articles with LLMs. Our findings revealed that the LLMs' recommendations reduced the burden on authors and enhanced understanding of the readers' perspective. However, our participants tended to rely on LLMs in the co-creation process, expecting them to generate useful questions and answers, leading to minimal edits and a passive role in content creation. Consequently, while QA-style articles created by our participants are easily comprehensible, they often fall short of showcasing the authors' distinctive voices and deep insights. From these insights, we propose future directions for interactions that support authors creating QA-style articles to promote their papers. The desirable direction for the interaction between humans and LLMs positions authors more engaging in the co-creation process, encouraging them to actively shape questions and critically monitor LLMs' recommendations.

ACKNOWLEDGMENTS

This work was supported by LG AI Research. We thank our participants for their engagement and the anonymous reviewers for their thoughtful comments and suggestions.

REFERENCES

[1] Heidi G Allen, Tasha R Stanton, Flavia Di Pietro, and G Lorimer Moseley. 2013. Social media release increases dissemination of original articles in the clinical pain sciences. *PloS one* 8, 7 (2013), e68914.

[2] Vera Septi Andrini. 2016. The Effectiveness of Inquiry Learning Method to Enhance Students' Learning Outcome: A Theoritical and Empirical Review. *Journal of Education and Practice* 7, 3 (2016), 38–42.

[3] Atlas.ti. accessed 2023. *Atlas.ti.* Retrieved September 14, 2023 from https://atlasti.com

[4] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 5327–5344.

[5] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* (apr 2023). https://doi.org/10.1145/3589955 Just Accepted.

[6] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. arXiv:2202.01279 [cs.LG]

[7] Dilan Bakır and Mehmet S Aktas. 2022. A Systematic Literature Review of Question Answering: Research Trends, Datasets, Methods. In *International Conference on Computational Science and Its Applications.* Springer, 47–62.

[8] S Elizabeth Bird and Robert W Dardenne. 2009. Rethinking news and myth as storytelling. In *The handbook of journalism studies.* Routledge, 225–237.

[9] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* 1–14.

[10] 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

[11] Susan M Brookhart. 2018. Appropriate criteria: Key to effective rubrics. In *Frontiers in Education*, Vol. 3. Frontiers Media SA, 22.

[12] EeeLN H Buckarma, Cornelius A Thiels, Becca L Gas, Daniel Cabrera, Juliane Bingener-Casey, and David R Farley. 2017. Influence of social media on the dissemination of a traditional surgical research article. *Journal of surgical education* 74, 1 (2017), 79–83.

[13] Xiang 'Anthony' Chen, Chien-Sheng Wu, Lidiya Murakhovs' ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. [n. d.]. Marvista: Exploring the Design of a Human-AI Collaborative News Reading Tool. *ACM Transactions on Computer-Human Interaction* ([n. d.]).

[14] Jean-Peïc Chou, Alexa Fay Siu, Nedim Lipka, Ryan Rossi, Franck Dernoncourt, and Maneesh Agrawala. 2023. TaleStream: Supporting Story Ideation with Trope Knowledge. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* 1–12.

[15] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[16] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–19.

[17] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology.* 1–13.

[18] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. OpenPrompt: An Open-source Framework for Prompt-learning. arXiv:2111.01998 [cs.CL]

[19] Dina El-Zanfaly, Yiwei Huang, and Yanwen Dong. 2023. Sand-in-the-loop: Investigating embodied co-creation for shared understandings of generative AI. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference.* 256–260.

[20] Elicit. accessed 2023. *Elicit.* Retrieved September 14, 2023 from https://elicit.org/

[21] Mireia Esplugas. 2023. The use of artificial intelligence (AI) to enhance academic communication, education and research: a balanced approach. *Journal of Hand Surgery (European Volume)* 48, 8 (2023), 819–822.

[22] Firebase. accessed 2023. *Firebase.* Retrieved September 14, 2023 from https://firebase.google.com/

[23] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference.* 1002–1019.

[24] Katy Ilonka Gero, Vivian Liu, Sarah Huang, Jennifer Lee, and Lydia B Chilton. 2021. What makes tweetorials tick: How experts communicate complex topics on twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

[25] Saloni Haldule, Samira Davalbhakta, Vishwesh Agarwal, Latika Gupta, and Vikas Agarwal. 2020. Post-publication promotion in rheumatology: a survey focusing on social media. *Rheumatology International* 40, 11 (2020), 1865–1872.

[26] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[27] Bilal Khalid Khalaf and Zuhana Bt Mohammed Zin. 2018. Traditional and inquiry-based learning pedagogy: A systematic critical review. *International Journal of Instruction* 11, 4 (2018), 545–564.

[28] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 115–135.

[29] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E Mackay. 2020. Imagesense: An intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.

[30] Kaisu Koivumäki, Timo Koivumäki, and Erkki Karvonen. 2020. "On social media science seems to be more human": exploring researchers as digital science communicators. (2020).

[31] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[32] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30 (2020), 121–204.

[33] Tomas Lawton, Kazjon Grace, and Francisco J Ibarrola. 2023. When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1978–1996.

[34] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.

[35] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. QASA: Advanced Question Answering on Scientific Articles. In *International Conference on Machine Learning*. PMLR.

[36] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse* 3, 2 (2012), 101–124.

[37] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[38] Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 241–252.

[39] Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynnedy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B Chilton. 2023. Tweetorial Hooks: Generative AI Tools to Motivate Science on Social Media. *arXiv preprint arXiv:2305.12265* (2023).

[40] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[41] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System That Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 454, 18 pages.

[42] Siobhán McHugh. 2007. The aerobic art of interviewing. *Asia Pacific Media Educator* 18 (2007), 147–154.

[43] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.

[44] Nature. accessed 2024. *Nature Podcast.* Retrieved Februrary 06, 2024 from https://www.nature.com/nature/articles?type=nature-podcast

[45] Eric Nichols, Leo Gao, and Randy Gomez. 2020. Collaborative storytelling with large-scale neural language models. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–10.

[46] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[48] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[49] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[50] Margus Pedaste, Mario Mäeots, Leo A Siiman, Ton De Jong, Siswa AN Van Riesen, Ellen T Kamp, Constantinos C Manoli, Zacharias C Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review* 14 (2015), 47–61.

[51] Perplexity.ai. accessed 2023. *Perplexity.ai.* Retrieved September 14, 2023 from https://www.perplexity.ai/

[52] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[53] Microsoft Research Podcast. accessed 2023. *Microsoft Research Podcast.* Retrieved September 14, 2023 from https://www.microsoft.com/en-us/research/podcast/

[54] Microsoft Research. accessed 2024. *Microsoft Research Podcast.* Retrieved Februrary 06, 2024 from https://www.microsoft.com/en-us/research/podcast/

[55] 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*. 1–13.

[56] Semantic Scholoar. accessed 2023. *Semantic Scholar.* Retrieved September 14, 2023 from https://www.semanticscholar.org/

[57] Science. accessed 2024. *Science Podcast.* Retrieved Februrary 06, 2024 from https://www.science.org/podcasts

[58] Scispace. accessed 2023. *Scispace.* Retrieved February 8m 2024 from https://scispace.com/

[59] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. 2023. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–57.

[60] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences* 32, 6 (2020), 635–646.

[61] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 43–52.

[62] Jakob Tholander and Martin Jonsson. 2023. Design ideation with ai-sketching, thinking and talking with Generative Machine Learning Models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1930–1940.

[63] Joanna Thornborrow. 2007. Narrative, opinion and situated argument in talk show discourse. *Journal of Pragmatics* 39, 8 (2007), 1436–1453.

[64] Milka Trajkova, Manoj Deshpande, Andrea Knowlton, Cassandra Monden, Duri Long, and Brian Magerko. 2023. AI Meets Holographic Pepper's Ghost: A Co-Creative Public Dance Experience. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*. 274–278.

[65] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[66] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[67] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *J. Amer. Statist. Assoc.* 114, 528 (2019), 1574–1596.

[68] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

[69] Spencer Williams, Ridley Jones, Katharina Reinecke, and Gary Hsieh. 2022. An HCI Research Agenda for Online Science Communication. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–22.

[70] Chengyuan Xu, Kuo-Chin Lien, and Tobias Höllerer. 2023. Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 350, 15 pages. https://doi.org/10.1145/3544548.3581282

[71] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. (2023).

[72] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[73] Pezhman Zare and Jayakaran Mukundan. 2015. The use of Socratic method as a teaching/learning tool to develop students' critical thinking: A review of literature. *Language in India* 15, 6 (2015), 256–265.

[74] Yu Zhang, Changyang He, Huanchen Wang, and Zhicong Lu. 2023. Understanding Communication Strategies and Viewer Engagement with Science Knowledge Videos on Bilibili. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[75] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

[76] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[77] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, Article 5, 30 pages. https://doi.org/10.1145/3586183.3606800

APPENDICES

A QUESTION GENERATION PROMPTS

Fundamental Prompt
I'm the podcast host and preparing an interview with the author to introduce the following paper. (...detail prompt from each question creation...) IMPORTANT: Please make the Question concise and short. Put questions in an array format without numbering. Paper title: { paper title } Abstract: { abstract } Instruction: { additional prompt from author }

Table 5: Fundamental Prompt for Question Recommendation.

General Question Prompt
I need to create key 5 questions to introduce the paper. The questions range from high-level questions like motivation, background, and future work or direction to low-level, detailed questions like experimental details and results.

Table 6: Prompt for General Question Recommendation.

Follow-up Question Prompt

I want to develop the following question and answer.
Question: { question }
Answer: { answer }
Please create a 3 follow-up questions.

Table 7: Prompt for Follow-up Question Recommendation.

Personalized Question Prompt

I need to create 5 questions to present the paper that elicit the author's perspective and view on the paper. Please Generate questions related to the paper in light of the author's publication list, topic, and agenda.
Author Information: { author information }
Publication List: { publication list }

Table 8: Prompt for Personalized Question Recommendation.

Pre-prompt
The following list is a single researcher's publication list. Please introduce the researcher's research agenda and domain. Rather than focusing on individual papers, address the overall trend and themes of the papers. Keep it brief and factual. Keep it concise with bullet style. The length should be within 2 paragraphs The publication list is sorted by newest to oldest. Publication List: { publication list }

Table 9: Pre-prompt for Author Information.

B EXAMPLES OF QA-STYLE ARTICLES

Questions	**Answers**	**Taxonomy**	**Generated by**
Summarize this paper in 50 words!	The paper suggests that functional near-infrared spectroscopy (fNIRS) might offer a surrogate marker of fMRI activation, which would broaden our understanding of various populations, including infants.	Others	Author
Explain the key methods of this paper in 50 words!	In this paper, the authors used machine learning models and data augmentation to predict well-validated fMRI markers of human cognition from multivariate patterns of fNIRS. They applied a neural data augmentation technique and four ML models to the fNIRS activation patterns to test their predictive capabilities.	Method	Author
Explain more about those 4 ML models that were used in this paper?	The four machine learning (ML) models used in this paper are linear regression, Lasso regression, ridge regression, and support vector regression (SVR) with radial basis function kernel.	System	Author
Which brain areas did the researchers focus for this study and why?	The researchers focused on the prefrontal cortex for this study because it is a non-invasive optical neuroimaging technique that measures hemodynamic responses in the brain using near-infrared light. This area is particularly important for human cognition and is often used in fMRI-based markers of individual differences.	Aim	Author
Explain the tasks that they used to investigate prefrontal cortex activation.	The tasks used to investigate prefrontal cortex activation in the study were the stop signal task and the probabilistic reversal learning task.	Method	Author
Using the 4 ML models and neural data augmentation, what did the researchers find?	Using traditional machine learning models, researchers were able to predict the true fMRI beta values by training their models on synthetic fNIRS beta datasets and fitting each estimated model to the true fNIRS beta values. The Lasso regression model with the HbR signals performed the best, predicting three out of eight clusters located in distinct regions of the brain.	Result	Author
What are these stop signal task and probabilistic reversal learning task?	The stop signal task and the probabilistic reversal learning task are two cognitive tasks used in the study to examine both low- and high-level cognitive abilities and their neural mechanisms.	Method	Author
Where do you see your research heading next? Are there other proposals to improve the predictive capacity or widen its utility for further studies?	The research is currently headed towards investigating whether fNIRS can predict functional connectivity examined through fMRI measurement with fNIRS. Additionally, investigating confounding variables such as environmental difference or emotional states might have affected the neural activation, which requires further investigation.	Application	General Question

Table 10: Example QA-style Article Created by P1.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Questions	Answers	Taxonomy	Generated by
What led you to tackle the problem of feature unlearning from a pretrained image generative model?	We develop a new framework to address the issue of feature unlearning in pre-trained image generative models in the image domain that is highly applicable to real-world situations. We propose using implicit user feedback to identify and remove target features from the model's latent representation, allowing for fine-tuning to exclude the production of samples with those features.	Opinion	General Question
Can you explain how your implicit feedback mechanism identifies which features to unlearn from the generated model?	The implicit feedback mechanism identifies which features to unlearn from the generated model by allowing users to select images that contain the target feature. Based on the feedback, a dataset with positive and negative examples is constructed. Once we obtain the latent vectors of the dataset, we use a vector arithmetic method to find the latent vector representing the target feature. We compute the mean vectors from a collection of positive images and negative images and subtract the mean vectors of the negative images from that of the positive images to get the target feature vector. With this target feature vector, we can identify the feature that user wants to erase.	System	General Question
Given your experiments on MNIST and CelebA datasets, can you discuss how successful the model was in removing target features while maintaining fidelity?	The results show that the unlearned model produces similar target feature ratios to the baseline for all features, indicating that the framework successfully unlearns the target feature. Additionally, all three models (Original model, unlearned model, baseline model) produced similar IS and FID scores, indicating that the framework can successfully unlearn the target feature while maintaining high-quality image generation.	Result	General Question
Given the results of your research, which additional applications do you see these methods contributing towards? What's the future work or direction in this domain from your perspective?	The unlearning algorithm for generative models has the potential to address concerns related to sensitive or private content, and there is scope for further research to enhance its effectiveness in other contexts like data privacy and fairness. Developing reliable unlearning algorithms can help maximize the benefits of generative models while minimizing risks.	Application	General Question
Your framework was tested on MNIST and CelebA datasets, what were these experimental findings or observations?	The experimental findings or observations of the framework tested on MNIST and CelebA datasets include the results of the Image Similarity (IS) and FID scores for evaluating the quality of generated images, respectively. The results demonstrate that all three models produced similar IS and FID scores, indicating that the framework can successfully unlearn the target feature while maintaining high-quality image generation.	Result	General Question

Table 11: Example QA-style Article Created by P7.