CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models

Riku Arakawa*
Carnegie Mellon University
Pittsburgh, USA
rarakawa@cs.cmu.edu

Hiromu Yakura*
University of Tsukuba / National
Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Japan
hiromu.yakura@aist.go.jp

Masataka Goto
National Institute of Advanced
Industrial Science and Technology
(AIST)
Tsukuba, Japan
m.goto@aist.go.jp

ABSTRACT

CatAlyst uses generative models to help workers' progress by influencing their task engagement instead of directly contributing to their task outputs. It prompts distracted workers to resume their tasks by generating a continuation of their work and presenting it as an intervention that is more context-aware than conventional (predetermined) feedback. The prompt can function by drawing their interest and lowering the hurdle for resumption even when the generated continuation is insufficient to substitute their work, while recent human-AI collaboration research aiming at work substitution depends on a stable high accuracy. This frees CatAlyst from domain-specific model-tuning and makes it applicable to various tasks. Our studies involving writing and slide-editing tasks demonstrated CatAlyst's effectiveness in helping workers swiftly resume tasks with a lowered cognitive load. The results suggest a new form of human-AI collaboration where large generative models publicly available but imperfect for each individual domain can contribute to workers' digital well-being.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; *Natural language interfaces*; • **Computing methodologies** → **Machine learning**.

KEYWORDS

large generative models, behavior change, task engagement, procrastination

*These authors contributed equally and are ordered alphabetically.

1 INTRODUCTION

Many office workers perform intellectual tasks, such as writing, coding, and editing slides, using computers every day. Supporting workers through the development of intelligent systems is an important topic that has long been studied in the field of human-computer interaction (HCI), and a number of tools have been proposed, such as efficient editors, for each task. An emerging trend in this context is to involve AI as a collaborative system. The recent rapid development of generative models has enabled replacing some portion of human tasks. For example, PEER [60] is a collaborative language model that can write drafts, add suggestions, and propose edits in writing. GitHub Copilot [86] is a pair programmer that can generate code corresponding to natural language descriptions of the target functionality.

While these systems that contribute to workers' tasks directly seem promising, their efficiency and subsequent workers' satisfaction depend on the accuracy and fidelity of the AIs. For instance, trust may not be built between systems and workers because of poor outputs that conflict with workers' intentions or do not meet their expectations [6, 26, 31, 81], resulting in ignorance of the systems. For example, Yang *et al.* [78] proposed an AI-based system that supports the writing of fictional stories through text generation; however, it was suggested that the output of the AI might irritate users due to its low accuracy, lack of usability, and insufficient transparency. It is theoretically possible to address this accuracy issue by developing more accurate models (*e.g.*, by fine-tuning large generative models in each target domain). However, it is not realistic to customize the models for each worker's task because of the huge cost of model training [3, 69] in a variety of tasks and domains.

In contrast to tuning systems specific to each task (*e.g.*, writing and slide-editing), we propose a new approach, *CatAlyst*, by designing AI systems that support workers' task engagement by inducing behavior changes using large generative models without tuning. During an intellectual task, workers often experience a lack of progress and stay away from it because they lose focus owing to the high cognitive load required to perform the task [43, 50, 73]. Such behavior often leads to procrastination, causing stress and degraded self-efficacy [8, 28, 34, 59]. We introduce a design employing generative models to influence workers' behaviors when faced with such moments by lowering the cognitive load necessary to resume working on their tasks, as *catalysts* in chemistry. Specifically, CatAlyst first detects when workers' progress is halted and then intervenes with them by presenting the continuation of their work,

which is generated by common large generative models. Importantly, CatAlyst aims to retrieve their interest and encourage them to resume the original task by lowering the hurdle, not to contribute directly to workers' tasks with generated content. We anticipate that current generative models can perform accurately enough to produce content to induce such behavior changes without any tuning for an individual task. Moreover, since workers recognize CatAlyst as an intervention to encourage their work rather than proactively using AIs with high expectations, they would not feel as irritated as systems like [78] when generative models perform poorly. The following is a scenario describing how CatAlyst is expected to work.

> Satoru is a university student who works on a presentation assignment. He is unmotivated and has put off working on it until the deadline of three days. He finished roughly 70% of the slide creation, but then no ideas came to mind, and he lost his concentration. He started watching popular drama on another tab of the browser. CatAlyst then detects his unengaged behavior and sends a notification on the browser, saying *"I generated the continuation of the slide. Let's take a look at it"*. He paid attention to the notification while watching the drama, became interested in the generated content, and glanced at the slide-editing interface. The content was thought-provoking, although the formatting and flow were a little off from what Satoru had put. Satoru began to ponder further the direction of the updated slide. He then stopped watching the drama and resumed the slide-editing task by formatting the automatically created slide and reorganizing the entire structure.

In terms of increasing workers' engagement via intervention, several approaches have been proposed in the HCI literature. For example, presenting an encouraging message or feedback is one useful approach in persuasive technology [11, 24, 40]. Previous studies have suggested that context awareness is important for effectiveness [56], and variation in the intervention content is no less essential for its continued effectiveness [33]. CatAlyst follows these studies, as it aims at presenting more context-aware and variational interventions as a form of continuation of interrupted work.

To test the effectiveness of CatAlyst, we conducted a series of evaluations. First, we developed a prototype running on a web page that supports writing tasks. It generates a continuation of sentences in progress using GPT-3 [15], a large language model. Our user study with 12 participants showed that CatAlyst's intervention has two significant effects compared with a conventional intervention using encouraging messages: (1) reducing the time the participants spent recovering their interest in the task and (2) increasing their productivity after the intervention. Furthermore, participants felt significantly less frustration while performing the task using CatAlyst and favored its usability.

Second, to examine how CatAlyst can be used in an unconstrained situation for a longer period, we conducted a user study in which ten participants used the prototype for five days. The participants used the prototype for different purposes, such as keeping

a diary, writing a novel as their hobby, and writing media articles as their job. Overall, their comments after the trial reaffirmed the usefulness of CatAlyst, demonstrating different cases in which our system helped them with writing tasks in their lives. Moreover, we found that they perceived CatAlyst's benefits from various perspectives, shedding light on the role of CatAlyst as a reminder, an ideator, and a peer.

Finally, to demonstrate the extensibility of CatAlyst in different tasks, we developed a second prototype in the form of a Chrome extension that supports slide-editing tasks. It generates the continuation of slides in progress using GPT-3 [15] by providing slides in progress as a few-shot prompt. It also incorporates a diffusion model [58] to generate an image that complements the generated slide. Our user study with 12 participants showed that CatAlyst's intervention had two significant effects compared to a conventional intervention: (1) reducing the time the participants spent recovering their interest in the task and (2) reducing the time they spent on the task by increasing their productivity. We also verified that CatAlyst posed less cognitive load to the participants, indicating the extensibility of our design.

While intelligent systems involving generative models tend to pursue their accuracy for each specific task to delegate larger parts of the tasks to AIs from humans, this is not always feasible owing to the issues with training data and computational costs. This study shows that, under the condition that large generative models have a certain degree of generality but not optimal accuracy, it is possible to support workers in multiple tasks by lowering the hurdle for resumption. Our results and discussion suggest a new form of human-AI collaboration that contributes to the digital well-being of office workers today.

2 RELATED WORK

This study is placed in the context of human-AI collaboration, as it aims to support knowledge workers with AI techniques. We first draw from the recent research trend of leveraging generative models to improve worker productivity while performing tasks. Thereafter, we review and emphasize the need to support workers' task engagement to highlight the difference between our approach and previous studies. Finally, we discuss factors affecting workers' task engagement based on existing HCI studies and introduce our idea of utilizing generative models for this purpose.

2.1 Generative Models for Improving Task Efficiency

As many tasks have been digitized, there is a strong need for computational support to improve the efficiency of such tasks, producing many software systems designed for various tasks, such as Microsoft Office 365, Adobe Creative Cloud, *etc.* Such systems have now started to incorporate AI techniques to add machine intelligence to accelerate workers' task performance. For example, Gmail [17] can automatically generate response candidates as well as complete sentences. Adobe Photoshop [7] features neural filters that allow users to easily manipulate, for instance, people's hair color, emotion, and light condition. GitHub Co-Pilot [86] can generate programming code based on the natural language description of the functionality provided by users. These products demonstrate

the enormous potential of leveraging AI techniques to directly contribute to and replace parts of human tasks, significantly improving workers' task efficiency.

More features and interactions have been proposed at the research level using machine-learning-based generation techniques. One of the most prominent examples is writing support systems that leverage large language models [15]. There are a number of recent studies that demonstrate how human writers and AI-based generation techniques can collaborate to support writers [23, 35, 60], such as in creative writing [57, 63, 65, 78, 82]. For example, Dang *et al.* [23] proposed a writing support tool that provides paragraph-wise summaries as self-annotations while workers write the content. They found that the generated summaries gave the workers an external perspective on their writing and helped to polish the content.

Emerging generation techniques have also been proposed for tasks other than writing, such as slide-editing [37, 62, 85], music creation [41, 70], and drawing [32, 38, 83, 84]. For example, Sefid *et al.* [62] proposed a slide generation technique from PDF files for scientific papers, whereas Zheng *et al.* [85] proposed an approach of converting computational notebooks into slides and showed that it improved the efficiency of a slide-editing task.

While these studies show promising results of employing AI-based generation techniques to improve task efficiency, there are two gaps: the wall of accuracy and the dependency on worker engagement. First, insufficient accuracy and fidelity in AI's output [67] can devastate trust in human-AI collaboration, leading to ignorance of such systems. Yang *et al.* [78] found that poor quality in generated texts could irritate fiction writers. Kim *et al.* [32] found that poor quality of output from AI does not provide rich inspiration for workers. It would be ideal to customize models for each worker so that they can achieve sufficient accuracy in individual tasks [66]; however, this is not necessarily feasible because of the huge cost of training and maintaining large models [3, 69]. Furthermore, Yang *et al.* [79] pointed out the existence of inherent uncertainty surrounding their capabilities and output complexity, making it difficult to construct trustful systems, even with such customized models. Given these inherent concerns, it is no less important to study generalizable approaches as to how generative models can contribute to workers performing various tasks without individual model tuning.

Second, the above studies implicitly assumed situations where workers are amenable to using the systems and continue working on tasks, which might not always hold in actual cases. In fact, as we describe in the next section, workers frequently struggle to maintain focus while performing tasks for various reasons, including the required high cognitive load and multitasking. When not only their attention is away from the task but also their motivation is low, workers might not fully benefit from the systems even if they feature more useful functions.

To overcome these two gaps, we propose a novel design of employing AI-based generation techniques to improve workers' task engagement without having to tune models to an individual task. Our approach targets workers who may not be fully immersed in their task, which is an alternative to previous studies that aim to improve the productivity of workers who already engage in the task.

We discuss related work on computational support for workers' task engagement in the following subsections.

2.2 Needs for Supporting Worker's Task Engagement

Modern knowledge workers need to concentrate on a task for a long time, such as writing, coding, and slide-editing. However, it is difficult to maintain concentration for a long time because these tasks cause mental fatigue; for example, human concentration is known to break after 30–45 min when performing these tasks [43, 50]. It is natural for people to become distracted, and this tendency is even higher in remote work scenarios [73], which have increased significantly in recent years owing to COVID-19 [51]. Furthermore, since such tasks require a high cognitive load, workers often experience difficulty, particularly when starting tasks or resuming after interruption. For example, Mark *et al.* [45] revealed that it could take 25 min to return to a task on average once distracted from the task. This further leads to procrastination [28, 34], which is known to be a major problem for office workers today, as it can lead to stress and problems in relationships, work, and health, such as depression [8, 18, 59]. Therefore, there is a crucial need to support workers dealing with these issues and increase their motivation and engagement in intellectual tasks.

Regarding workers' task engagement, it is known that the causes of the high hurdle to starting work are complex [74]. One of the major causes is perfectionism among workers, leading to their assumption that tasks must be completed flawlessly. They often stop thinking before starting tasks because they consider too much about the expected workload and sometimes feel fear of failure [61]. In response, practitioners often emphasize the importance of starting tasks anyway, for instance, for 3–5 min before deeply thinking about the workload [53], as one of the best ways to engage in the tasks. The Pomodoro technique [20] is a strategy to divide the entire task into small segments so that workers can start working on one of them with a lowered hurdle.

Our strategy is to design intervention by leveraging generative models to induce such behavior change, that is, making it easier for workers to take the first step to start working when they are not immersed in the task. In the next subsection, we review existing work in the same context as supporting workers' task engagement and discuss how generative models could be a powerful approach to achieving it.

2.3 Interventions for Supporting Worker's Task Engagement

There are several approaches proposed to support workers' task engagement through interventions when they are distracted. Site blockers are a typical example of explicit intervention since they force users to return to their intended task by blocking access to specific entertainment websites. However, this is a primitive and static approach that increases the psychological burden on the user owing to its coercive power [33], resulting in workers uninstalling the systems.

By contrast, several studies in HCI also design interventions that affect people's behavior less explicitly. Arakawa and Yakura [5]

proposed *mindless intervention* to draw students' attention to online classes based on the design of Mindless Computing [1]. Since such mindless interventions are designed for situations that do not demand users' conscious awareness, they could be effective in our case to draw workers' attention to the task (*e.g.*, writing). However, this assumption is not always applicable when we try to get users to actually begin the task, given the higher hurdle imposed, as discussed in relation to procrastination.

Nudge has been extensively studied [11, 12, 24, 40, 44, 56] as a bit more explicit intervention than mindless approaches. For example, encouraging messages have been popularly employed to motivate students [11] or induce exercise behavior [24]. Furthermore, Liu *et al.* [40] proposed a method to motivate workers to return to the task of editing ill-formatted documents by designing feedback reflecting their progress based on the perspective of persuasive technology. Their user study showed a significantly shorter average off-task time by adding visually encouraging feedback that reflected their progress. Similarly, Rodriguez *et al.* [56] suggested presenting context-aware content as a factor for successful nudge design, as it improves personal relevance and motivation. Even though such nudging approaches can also be effective in our case, the inherent nature of nudging would not always be suitable in the context of procrastination on intellectually-demanding tasks. Nudge has emphasized the importance of providing all available choices to people, without forbidding any of them [71], from the perspective of libertarian paternalism. Meanwhile, people who procrastinate intentionally choose to do so while recognizing the necessity of completing the tasks [28, 34]. Given this, the effectiveness of such nudge-based approaches is uncertain for our situation, where the target tasks would require higher effort and focus than those in prior studies, such as editing ill-formatted documents.

Thus, to make these nudge-based approaches more sophisticated, we introduce an approach of providing the continuation of interrupted work to lower the hurdle for resuming a task. As in the example of Satoru described in Section 1, workers prompted with imperfect but relevant work progress would try to work on it for 3–5 min instantly, which is important to avoid workers' perfectionism and procrastination behavior. We leverage generative models for this purpose since recent large generative models are adept at producing context-aware content, although they are sometimes unable to satisfy workers who expect the generated content to replace parts of their tasks directly [32, 78]. This point guided us to anticipate that such generated content could be used to support workers' task engagement by drawing their diverted attention and offering a base to work on. More specifically, we expected that workers can resume the interrupted task with less effort if they find the generated content supporting their ideation or providing relevant information.

3 PROPOSED APPROACH: CATALYST

In this section, we describe our approach of intervening with workers using the generated continuation of interrupted work. We also elaborate on the strategy behind this idea from the perspective of persuasive technology.

3.1 Description

3.1.1 Overview. Figure 1 shows an overview of our approach of intervening with workers to enhance task engagement. CatAlyst first detects the moments where the progress of the workers' tasks is stopped from the interaction log (A). It then generates content that is likely to follow interrupted work using large generative models (B). For example, GPT-3 [15] can be used in writing tasks to predict what would come next. Finally, it sends notifications to the workers while showing part of the generated content to induce them to resume tasks (C). To expand the applicability, we designed the pipeline so that we can use any generative model as long as it can produce a context-aware continuation of the interrupted task. In this study, we implemented two prototypes: a writing support tool and a slide-editing support tool. Both prototypes run as web applications.

3.1.2 A: Detecting interruption. Previous research has explored several approaches for detecting a break in a worker's concentration, including image-based sensing using a webcam, as in [5], and estimation from interaction log [40]. In this study, we adopted the method of estimating using interaction logs to avoid the additional cost and risk of using sensors (*e.g.*, privacy). Specifically, our prototypes send the interaction logs to a server and find a moment when the workers have not made any interaction for T s. The initial value of T was determined in a pilot study as the approximate time that distinguishes the actual distraction of the workers from the random interval during an engagement, which was 45 s. We were aware that, even if workers do not make any interaction over T s, they might engage in tasks, such as long contemplation, resulting in false-positive detection. Simultaneously, we presumed that intervening with them in such moments using the generated continuation would not work negatively, as it could provide them with new perspectives or information that could facilitate their contemplation. This idea of designing an intervention that might work (or at least does not cause negative experiences) even when triggered by false-positive detection was inspired by Mindless Attractor [5]. Note that, as we received feedback from workers in our experiments, we can make the value T controllable to provide them with global control of AI systems for a better human-AI relationship [2].

3.1.3 B: Generating continuation of interrupted work. We applied large generative models, such as GPT-3 [15], to generate the continuation of the interrupted work. In the writing task, for instance, we provided the text in progress to a model for predicting the text to follow. The same scheme can be applied to the slide-editing task; we extracted all text fragments and image captions from the slides in progress and provided them to a model (Figure 2). Subsequently, the continuation prediction from the model is used to fill the new slide that is automatically appended. Here, CatAlyst can be applied to other domains by alternating the model to be used with, for example, those for music creation [41, 70] and drawing [32, 38, 83, 84]. Furthermore, the slide-editing task illustrates how our approach of harnessing large generative models can be extensively applied to new tasks by formulating appropriate inputs (*i.e.*, prompt programming [55]) even if those tasks have not been considered at the time of training.
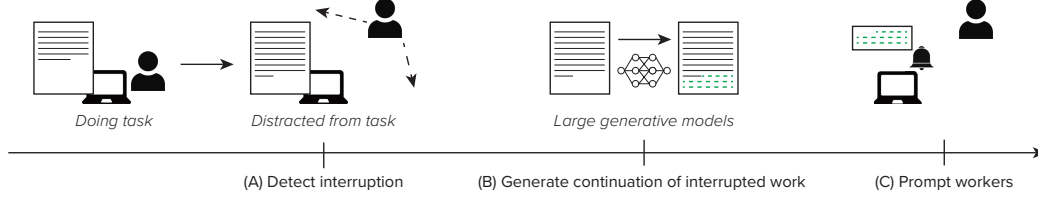
Figure 1: Overview of CatAlyst. (A) Detecting interruption based on workers' progress. (B) Generating continuation of interrupted work using large generative models. (C) Prompting workers to resume the task via notification.
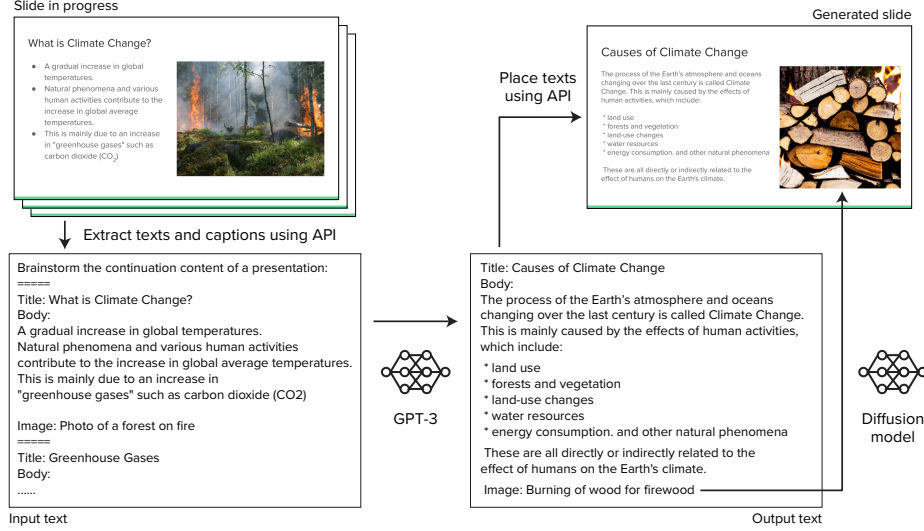

Figure 2: Actual example of how CatAlyst can generate the continuation in the slide-editing task using GPT-3.

3.1.4 C: Prompting workers. Our prototypes notify workers using a Notification API of the browsers once the continuation is generated. We aimed to draw not only workers' attention but also their interest, as discussed later in Section 3.2.1. Therefore, we designed the notification to present the beginning of the generated continuations (*e.g.*, the first sentence of the generated text or the title of the generated slide). We note that CatAlyst can be extended to send notifications in other forms, such as using a smartphone to stimulate workers whose interest is deprived by smartphone apps.

3.2 Strategy

Resuming an intellectual task while a worker gets distracted is expected to require more cognitive load than in cases where a nudge or mindless intervention is successful, such as getting students' attention back [5]. In these situations, we expected that encouraging feedback, as in conventional methods [24, 40], has room for improvement to provide more persuasive interventions and that prompting workers along with a continuation would be more effective. By doing so, CatAlyst encourages workers to start thinking about the task regardless of their status, even for a short time, which is known as a popular practice to engage in the task [53]. This idea was originally inspired by a practitioner's insight, which has been popularly adopted by readers. James Clear [21] proposed a strategy

for behavior change by employing an analogy of *activation energy* in chemistry. He explained that every habit has activation energy that is required to get started, and thus, it is important to introduce a *catalyst* to lower the activation energy. He further suggested utilizing intermediate steps with lower activation energy to induce behavior changes step by step. Our strategy of achieving such a *catalyst*-like effect using the generated continuation of interrupted work as the intervention can be broken down in detail, as shown in Figure 3.

3.2.1 Guiding workers' interest to generated contents. First, it is important that workers become interested in interventions and do not ignore them. Regarding this point, interventions via notification can at least draw workers' attention but do not always draw their interest. This means that it is easy to anticipate that workers will ignore them if the content of the intervention is not very useful [68] or attractive, as we often ignore push notifications on smartphones (the first failure case in Figure 3). Here, Kovacs *et al.* [33] showed that adding variety to intervention is effective in maintaining novelty. According to Berlyne [10], inducing curiosity is important to drive motivation. Wickens [75] emphasizes the importance of inherent interest based on a computational model on top of cognitive science.
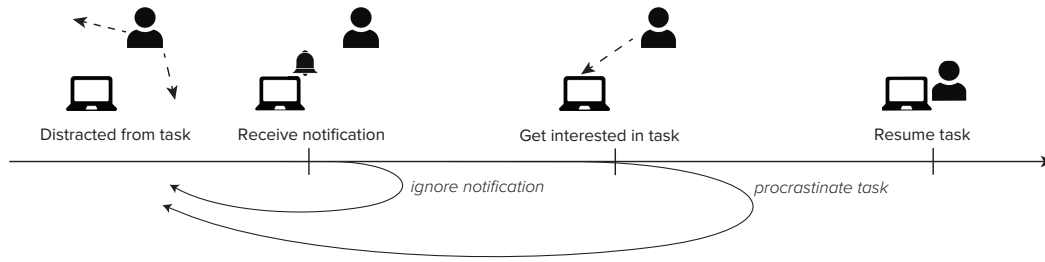
Figure 3: Interventions lead to workers' task resumption through multiple steps. At each step, there is a chance for an intervention to fail to induce workers' behavior change.

Our strategy is to make workers interested in the content of the intervention by presenting the generated continuation of their interrupted work. For example, in the case of chatbots, people's curiosity about AI's outputs is reported as one of the major factors for them to use the technology [13, 39]. In this regard, CatAlyst generates continuation of interrupted work, which is more context-aware information, leading to the inherent interests of workers.

3.2.2 Inducing workers' task resumption. The second part of the intervention flow (Figure 3) is to induce workers' behavior of resuming the task after they retrieve their interest in the original task. More specifically, our objective is to prompt them to resume the task quickly for 3–5 min to avoid procrastination and engage in it. According to Fogg's behavior model [27], a behavior change is achieved when the elements of motivation, ability, and prompts are aligned. Regarding motivation, if workers perceive CatAlyst as a collaborator, they will gain motivation to utilize the generated content, as demonstrated in prior studies on human-AI collaboration [83, 84]. Regarding ability, the generated continuation of the interrupted work can be a help, for example, by supporting their ideation or providing reference information, as in prior studies about human-AI co-creation reviewed in Section 2.1. Therefore, we expect CatAlyst's design to guide workers to resume the original task with enhanced motivation and ability, preventing procrastination.

3.3 Hypothesis

The design strategy of CatAlyst, as discussed thus far, poses the following hypotheses. First, we should examine whether workers can recover their interest by being presented with the generated content as an intervention, as this is the first step toward resumption (see Figure 3).

> H1: CatAlyst is an effective means to keep attracting the interest of workers who are away from the task by presenting the continuation of interrupted work as an intervention.

Next, even if the workers' interest returns to the task by seeing the generated continuation, they might not resume the task and go back to the distracted states, *i.e.*, procrastination, given the required hurdle for working on the task. Therefore, we need to further investigate whether the workers prompted by the intervention

change their behavior (*i.e.*, resume the task), guiding us to posit the following hypothesis.

> H2: CatAlyst can induce workers' behavior to resume the original task effectively through the intervention.

If these hypotheses are supported, the effectiveness of CatAlyst in supporting workers' task engagement will be validated. We further anticipate that the overall productivity of workers would be enhanced with the intervention, leading to the following hypothesis. If this hypothesis is supported, CatAlyst can be used as an alternative approach to the previous studies to improve task efficiency.

> H3: CatAlyst can improve worker productivity by helping them avoid procrastination while performing tasks.

Finally, we expect that CatAlyst can contribute to improving the subjective experience of workers who often face a high cognitive load during tasks. If this hypothesis holds true along with the above hypotheses, we can conclude that our strategy behind the design of CatAlyst functioned as intended. Therefore, we examine whether CatAlyst actually lowers the hurdle for workers to resume tasks and offers a favorable experience.

> H4: CatAlyst can lower the cognitive load imposed on workers while performing a task, thereby being favorably accepted by them.

4 STUDY 1: WRITING TASK

To test our hypotheses, we first conducted a user study in a controlled setting in which we prepared a prototype of CatAlyst for supporting workers' writing. This is because, while writing is one of the tasks that knowledge workers regularly perform, it requires a high level of skills and cognitive load, and workers often lose focus while performing it [46].

4.1 Design

We conducted a within-participant study to compare three conditions: *proposed*, *control*, and *none*. Participants experienced interventions from our prototype of CatAlyst in the *proposed* condition. In the *control* condition, they instead experienced a conventional intervention for supporting task engagement, that is, encouraging

messages (see Section 2.3). In the *none* condition, they did not experience any intervention and only performed the writing task as usual.

4.2 Task

In this study, we had participants write essays of approximately 1,200 characters in Japanese. We prepared six essay tasks based on publicized university exams to balance their difficulty, such as: "There are more people who eat alone these days. What do you think about this trend from the perspective of modern social structure?" Three were randomly chosen for each participant and assigned to the three conditions. It took roughly 30 min to write one essay.

4.3 Implementation

We prepared three types of web pages corresponding to the three conditions. As shown in Figure 4, they had a center box in which the participants wrote texts. In the *proposed* and *control* conditions, the web pages recorded the interaction log of the participants and detected the moment they got distracted. In the *proposed* condition, the written text in progress was sent to a remote server, where its continuation was generated using GPT-3 [15] pretrained with the C4 [54] and CC100 [22] datasets[1]. The web page prompted the participants with a notification displaying its first sentence once the continuation was generated. In the *control* condition, the web page did not send the text and immediately showed a notification displaying an encouraging message that was randomly chosen from the six patterns prepared based on a public site blocker[2].

4.4 Procedure

We published a web page for the experiment and recruited participants and, to secure the number of participants, we posted a link to the page on social media. As a result, twelve participants (six males and six females; 24–59 years old) participated in the study. Each two of them were assigned to one of the six possible orders of the three conditions. The participants agreed to the research policy and were provided with a brief explanation of the usage of the web pages in the beginning. Here, we clarified that we do not penalize them even if they spend a long time on the tasks so that they can perform the tasks as usual. In addition, we told them that they were allowed to perform any habit they might have in performing tasks, for example, listening to music, to induce spontaneous behavior, following the procedure of Liu *et al.* [40]. They then performed three essay tasks using the web page that corresponded to the assigned order. Every time they completed a task, they were asked to fill out a questionnaire to evaluate their experience of the task using the web page. They were also asked to fill out another questionnaire that requested a comparison of the three conditions and additional comments after completing all the assigned tasks. The study was conducted remotely for all participants.

---

[1]https://huggingface.co/rinna/japanese-gpt-1b
[2]https://chrome.google.com/webstore/detail/blocksite-block-websites/eiimnmioipafcokbfikbljfdeojpcgbh

4.5 Measure

We used multiple measures according to our hypotheses. From the participants' behavior during the use of web pages, we calculated four measures: interest retrieval time, ignorance rate, progress after resumption, and total writing time. From the questionnaire responses, we evaluated cognitive load and system usability.

4.5.1 Interest retrieval time. In correspondence with H1, we calculated *interest retrieval time*, the time each participant spent from the moment they received a notification to the moment they recovered their interest in the task. Here, we caught the recovery of interest by finding the moment when they made any operations on the web page (*e.g.*, moving a mouse or pressing a key) for the first time after receiving a notification. Furthermore, we compared the values of the *proposed* and *control* conditions because the *none* condition did not present any notifications. When the interest retrieval time of the *proposed* condition was shorter than that of the *control* condition, H1 is supported.

4.5.2 Ignorance rate. We also calculated *ignorance rate*, a rate of interventions (*i.e.*, notifications) that were ignored by the participants. The ignorance rate is highly correlated with the interest retrieval time (*i.e.*, a longer interest retrieval time leads to a higher ignorance rate) because the *proposed* and *control* conditions prompted a notification every T s until the participants made any operations. Still, we prepared this measure to evaluate H1 precisely. Specifically, as discussed in Section 2.1, there was a risk of losing the trust of the participants due to a mismatch with their expectations, fostering ignorance of the notification. Therefore, to confirm that CatAlyst can keep attracting their interest, we examined how the ignorance rate changed as the number of interventions each participant received increased. If the ignorance rate of the *proposed* condition was lower than that of the *control* condition and exhibited stable values with respect to the increase in the number of interventions, H1 is further supported.

4.5.3 Progress after resumption. In correspondence with H2, we quantified and compared the number of characters the participants typed within a specific period after the participants recovered their interest in the task, namely, *progress after resumption*. As discussed in Section 3.2.2, they might procrastinate on the task even though they have recovered their interest. In such cases, progress after resumption would remain insignificant. Conversely, H2 is supported if the increase in the number of characters after resumption in the *proposed* condition was more significant than that in the *control* condition.

4.5.4 Total writing time. We evaluated the productivity of the participants via two measures. The first is the *total writing time* they spent on the tasks. The average of the total writing time would be equivalent if CatAlyst did not contribute to the improvement of worker productivity since we balanced the difficulty and assignments of the tasks prepared. In other words, H3 is not supported unless the total writing time of the *proposed* condition was shorter than that of the other two conditions.

4.5.5 Subjective quality. We also evaluated the *subjective quality* of the texts the participants wrote. This was to ensure that the reduction in the total writing time was enabled not by sacrificing the text

(a) The participants performed the tasks on this interface.

(b) The continuation of their work was generated and notification was sent.

(c) They were able to refer to the generated continuation in the *proposed* condition.
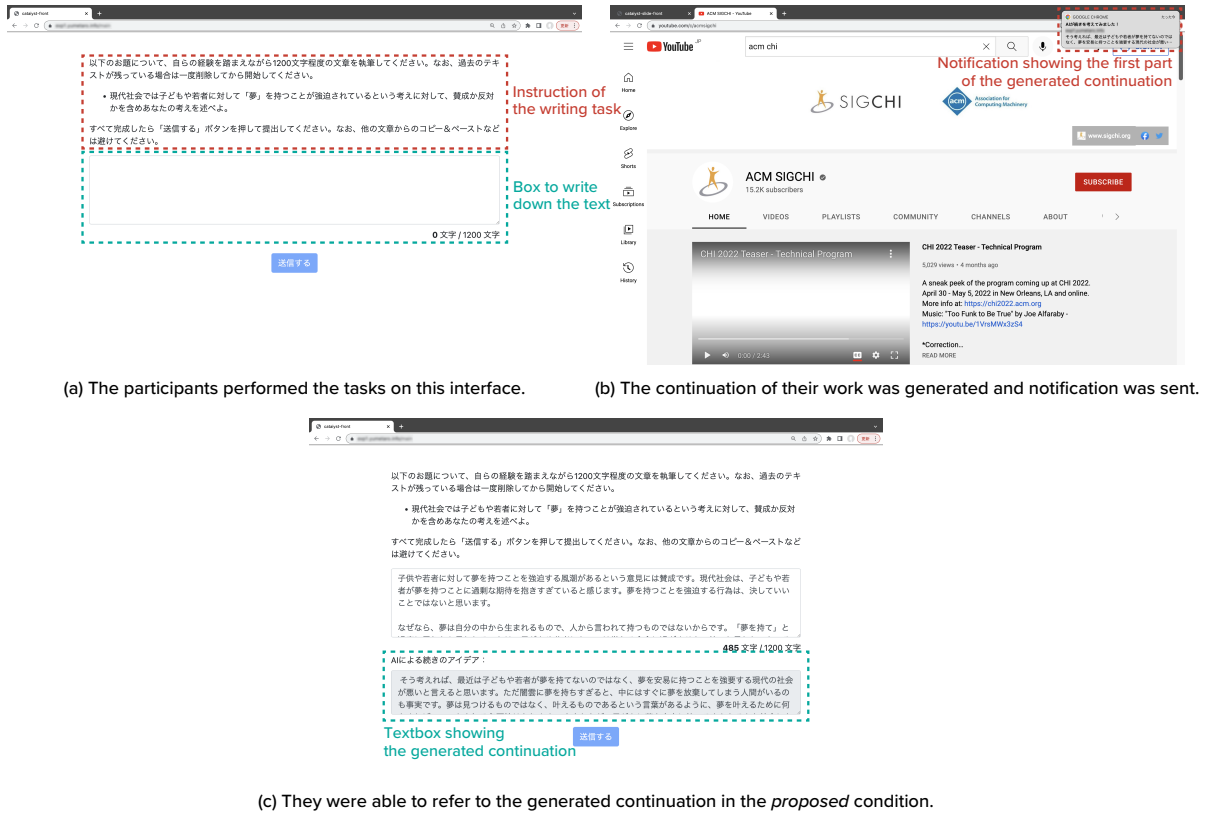
Figure 4: Interface used in Study 1.

quality, but by the improvement in their productivity. Therefore, we asked three volunteer raters who had not participated in the experiment to rate the texts in a manner agnostic to the experimental conditions. Each text was evaluated for three aspects using a 7-point Likert scale: consistency, readability, and overall quality. Therefore, H3 is supported when the *proposed* condition shortened the total writing time while the raters' evaluation for the texts written in this condition, at least, did not deteriorate compared to the other two conditions.

4.5.6 Cognitive load. We then examined H4 using the questionnaire. Specifically, we used the six items of the NASA-TLX [29] and calculated the raw TLX score [16]. This allowed us to examine *cognitive load* across six factors: mental demand, physical demand, temporal demand, performance, effort, and frustration. As higher scores indicate higher cognitive load, H4 is supported when the scores of the *proposed* condition were lower than those of the other conditions.

4.5.7 System usability. Furthermore, we examined *system usability* to complement the cognitive load in terms of evaluating the subjective experience of the participants. We adopted ten items from the System Usability Scale (SUS) [14] and compared their usability evaluation for the *proposed* and *control* conditions. The case in which the score of the *proposed* condition was higher than that of the other conditions corroborates H4 because it implies

better usability of CatAlyst. Moreover, we asked the participants to rank the three conditions in order of their evaluation of system usability in the questionnaire they completed at the end of the study. We compared their responses across the three conditions to confirm that the *proposed* condition was ranked higher than the other conditions, which leads to supporting H4.

4.6 Results

4.6.1 Interest retrieval time. First, we calculated the interest retrieval time, as presented in Table 1. The participants in the *proposed* condition reacted to the intervention within a much shorter period than those in the *control* condition. Welch's t-test after Levene's test confirmed that the mean of the values in the *proposed* condition was significantly shorter ($t(72.31) = 3.20, p = 0.002$) than that in the *control* condition, supporting H1. Furthermore, the *control* condition exhibited a greater variance in the interest retrieval time. In fact, we found cases where the encouraging message did not work for some completely distracted participants, whose interests were recovered after more than 10 min, conforming to previous research [45].

4.6.2 Ignorance rate. Moreover, H1 was corroborated by a comparison of ignorance rates. Corresponding to the longer interest retrieval time in the *control* condition, it presented a much higher

Table 1: Comparison of the interest retrieval time between the *proposed* and *control* conditions, which were significantly different ($p = 0.002$).

	mean		SD
control	135.0	±	303.7 (s)
proposed	18.8	±	25.1 (s)

Table 2: Comparison of the ignorance rate between the *proposed* and *control* conditions. While the rate was significantly lower for the proposed condition ($p < 0.001$), no difference was observed when we focused on only the first intervention presented to each participant ($p = 0.260$).[3]

	(A) **All**		(B) **First time**	
	ignored	worked	ignored	worked
control	185	71	4	11
proposed	5	29	0	8

number of notifications until the participants recovered their interest in the task, most of which were ignored, as shown in Table 2 (A). The ignorance rate of the *proposed* condition was significantly lower ($p < 0.001$) than that of the *control* condition, according to Fisher's exact test. Interestingly, we could not find a difference between the two conditions when we focused only on the first intervention presented to each participant ($p = 0.260$). In other words, the intervention of the *control* condition attracted the participants' interest at the beginning; however, as they saw more encouraging messages, they became less persuaded by the encouraging messages, leading to an increase in the ignorance rate. Conversely, this suggests that CatAlyst can keep attracting the interest of workers, which we intended by introducing the personalized, context-aware, and variational intervention.

In addition, this result can be explained by *perceived utility*, which Ling *et al.* [39] mentioned as a factor in maintaining the use of AIs. The workers will keep interested in the intervention if they feel usefulness in the generated content (*e.g.*, as support for ideation), whereas conventional encouraging interventions do not have such utility. We later discuss this point based on the *system usability* and comments of the participants.

4.6.3 Progress after resumption. We then examined the progress after resumption, as presented in Table 3. We found that the participants in the *proposed* condition made more progress within T s after their interest returned to the task than those in the *control* condition, according to the t-test ($t(98) = 2.08, p = 0.041$). This implied two possibilities regarding the participants' behavior: 1) participants in the *proposed* condition could utilize the generated content to come up with what to write next and 2) participants in the *control* condition procrastinated the task even after being guided to perform some operation on the web page by encouraging messages. This result supported H2.

4.6.4 Total writing time. As shown in Table 4, we could not find a significant difference in the total time the participants spent between the three conditions, according to the Friedman test after Levene's test. Thus, H3 was not supported. However, we can see that the participants in the *proposed* condition spent less time on average than those in the other conditions, which implies the effect of CatAlyst. Simultaneously, the average time spent in the *control* condition was slightly longer than that in the *none* condition. This can be attributed to the fact that the previous work in Section 2.3 employed a clear criterion for detecting interruptions, such as being presented with explicit interruption cues by an experimenter [40],

whereas we used time-based naïve criteria to examine the effectiveness of CatAlyst against false-positive detection.

4.6.5 Subjective quality. Meanwhile, we also could not find a significant difference in the raters' subjective evaluation of the quality of the texts the participants wrote between the three conditions, according to the Friedman test. Specifically, as shown in Table 5, all of the three aspects the raters evaluated (*i.e.*, consistency, readability, and overall quality) did not exhibit a significant difference ($p = 0.354, 0.221,$ and 0.972, respectively). This result implied that CatAlyst did not deteriorate the quality of the texts the participants wrote, while H3 was not supported.

4.6.6 Cognitive load. To evaluate H4, we also examined the participants' evaluations of their cognitive load, as shown in Figure 5. One-way repeated measures ANOVA suggested significant differences in their cognitive load with respect to the frustration factor ($F(2, 22) = 9.37, p = 0.001$). We then found that the frustration they felt during the *proposed* condition was significantly lower than those of the *control* ($t(11) = 3.36, p = 0.019$) and *none* ($t(11) = 3.25, p = 0.019$) conditions, according to the post-hoc test with Holm correction. This result suggests the plausibility of H4.

4.6.7 System usability. Finally, we examined the participants' evaluations of the system usability. As presented in Table 6, the scores obtained using SUS implied that the participants preferred the *proposed* condition over the *control* condition, while their difference was not significant according to the paired t-test ($t(11) = 1.91, p = 0.082$). At the same time, the participants' rankings regarding usability suggested that the three conditions made a significant difference in their rank order, according to the Friedman test ($\chi^2(2) = 8.91, p = 0.012$). Specifically, the post-hoc Nemenyi test [49] revealed that the *proposed* condition was likely to be ranked significantly higher (*i.e.*, regarded as having better usability) than the *none* condition ($p = 0.022$). In particular, we found that eight participants responded that the *proposed* condition was the best. These results indicated that CatAlyst was favorably accepted by participants, corroborating H4. This also implied that the *proposed* condition contributed to the perceived utility of the participants, and thus, its effectiveness did not deteriorate after repeated exposure, as mentioned above.

4.7 User Comments

From the comments the participants left, we can confirm that CatAlyst functioned as we expected, as follows.

---

[3]Total number of the notification presented to the participants as the first intervention in the *control* condition exceeds the number of participants because some of them did not recover their interest until several notifications were prompted even in the first intervention.
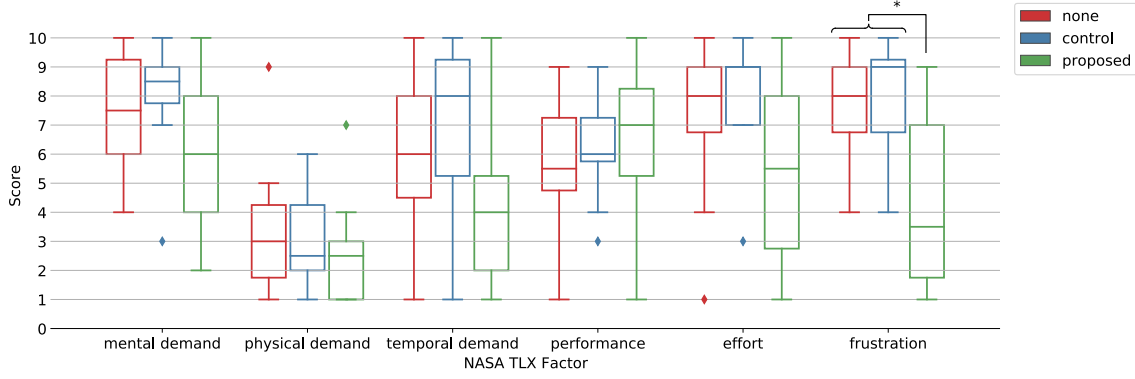
Figure 5: Participants' evaluations of their cognitive load in Study 1. The frustration they felt in the *proposed* condition was significantly lower than those of the *control* ($p = 0.019$) and *none* ($p = 0.019$) conditions.

Table 3: Comparison of the progress after resumption time between the *proposed* and *control* conditions, which were significantly different ($p = 0.041$).

	mean	SD
control	17.7 $\pm$	36.3 (chars)
proposed	63.5 $\pm$	186.2 (chars)

Table 4: Comparison of the total writing time between the three conditions, which were not significantly different ($p = 0.174$).

	mean	SD
control	2557.3 $\pm$	2196.5 (s)
proposed	1747.3 $\pm$	1660.0 (s)
none	2231.2 $\pm$	1359.8 (s)

> Not only did the notification make me realize that I was distracted, but it also helped me get back to the task as I wondered what the AI had written.

> I often lost concentration and stopped working because I could not come up with the next sentence, so it was easier to get back to work with being presented with the AI's text.

Furthermore, we found that the generated continuation helped their work by providing ideas and relevant information.

> I was very surprised the first time when the text from the AI was presented. It actually fit the topic, and I was able to use that as a reference for my writing.

> Although I did not find consistency or uniformity in the AI's output, I found it useful for imitating its writing style and obtaining ideas for paraphrasing.

The latter comment suggested that the participant positively perceived the use of CatAlyst even though the performance of the large language model was not completely satisfactory. Importantly, such a gap in performance expectations can lead to devastating trust, as discussed in Section 2.1. However, our design reframes the main goal of CatAlyst to guide participants' interest instead of directly contributing to their tasks. We suspect that this induced the participants' tolerance of the AI's performance and the enhancement of their subjective usability.

Our observations are further supported by examining their comments on the *control* condition, as follows:

> At first, it was a good trigger for recognizing that I was not concentrating. But, when I received the notification from the system while I interrupted the

> task to write a quick reply to an incoming message, I felt a little annoyed.

This comment not only emphasizes the importance of designing an intervention that also works with false-positive detection but also suggests the reason why the intervention of the control condition tended to be ignored as the number of interventions increased (Table 2).

We note that one participant requested that the frequency of the notification be controllable. This conforms to the previous study that emphasized the importance of providing global control in AI-based systems [2]. This comment helped us improve the prototype we provided in the following long-term study.

5 STUDY 2: WRITING TASK IN THE WILD

Thus far, the effectiveness of CatAlyst's design in the writing task has been validated by confirming H1, H2, and H4 through the controlled user study. Still, the results may be attributed to the fact that it might be the first time that the participants interacted with the recent large generative models. In other words, conducting a long-term study to deny the possibility that the effectiveness of CatAlyst depends on freshness is desirable. Therefore, we evaluated the effectiveness of CatAlyst in a study with a longer duration and participants having diverse motivations and purposes for writing. Ten participants used the prototype for five days and provided their perspectives and feedback on the design of CatAlyst.

Table 5: Comparison of the subjective quality of the texts between the three conditions, which were not significantly different regarding consistency, readability, and overall quality ($p = 0.354$, 0.221, and 0.972, respectively).

	consistency	readability	overall quality
control	5.50 ± 1.21	5.25 ± 1.32	5.11 ± 1.37
proposed	5.03 ± 1.61	5.17 ± 1.63	4.89 ± 1.65
none	5.47 ± 1.54	4.97 ± 1.46	5.06 ± 1.43

Table 6: Comparison of the participants' evaluations of the system usability between the *proposed* and *control* conditions, which were not significantly different ($p = 0.082$).

	mean	SD
control	62.1	± 22.8
proposed	75.6	± 13.7

5.1 Implementation

We reused the prototype used in Section 4. We enabled participants to adjust the time threshold T, which would be helpful, particularly when the prototype is used in the wild, based on the feedback we had gained in the previous study. Furthermore, we assumed that participants would leave the web page for a substantial amount of time because of other tasks in their daily lives (*e.g.*, having lunch) in contrast to the controlled setting in the first study. We thus added a feature of sending participants an email with generated content 5 min after they left the web page. We expected that this feature would help participants to resume the original task when they finished other tasks. Note that the content written by participants was neither accessible from experimenters nor stored on the server to preserve their privacy; it was only used for AI's inference and immediately discarded after it.

5.2 Procedure

In the same manner as in Study 1, ten participants (P1–P10) were recruited. Table 7 summarizes their backgrounds. They self-reported their writing experience regarding the period they regularly had writing tasks. In this study, the participants first agreed to the research policy and familiarized themselves with how to use the prototype. We asked them to use the prototype for writing freely during the study period whenever they would like, without setting any quota. After a participant spent five days, we asked for their comments to understand how they perceived the use of CatAlyst in terms of its usability and influence on their behavior. The participants were asked a series of questions: "*what kind of writing did you use the system for?*"; "*can you tell us your overall impressions of using the system?*"; "*did you feel any change with and without the system?*"; "*what is your impression about the AI which generated content?*"; "*what was this system like for you?*"; "*what would you like to see with this system in the future?*"; "*do you want to continue using the system?*" Note that we conducted the experiment remotely for all participants.

After the above procedure, we also examined the participants' use of CatAlyst from the interaction log. This aimed to confirm the long-term effectiveness of CatAlyst in a manner disassociated from the possible effect of the freshness in interacting with large generative models. Specifically, we calculated the *interest retrieval time* in the same manner as in the previous study and analyzed its trend over the five days of their use.

5.3 Usage results

As presented in Table 8, we could not find a significant difference in the interest retrieval time across the five days, according to one-way ANOVA after Levene's test ($F_{(4, 112)} = 0.425$, $p = 0.791$). This suggested that the effectiveness of CatAlyst in guiding the participants' interests did not disappear even after long exposure to it. We note that the higher values in Table 8 compared to the previous experiment (Table 1) are attributed to the difference in the situations. Specifically, these values include the moments when the participants interrupted their writing tasks for necessary reasons in their daily lives, while we asked the participants in Section 4 to complete the assigned tasks as a part of the experiment.

In addition, Table 8 allows us to infer that the participants continued the use of CatAlyst over the five days, whereas the occasions they used can be influenced by various factors (*e.g.*, weekdays or weekends) in the uncontrolled setting of this experiment. Note, since we did not store the content written by the participants, the progress they made cannot be quantified. Therefore, we then analyzed their responses in the semi-structured interviews we conducted.

5.4 Interview Results

Overall, participants responded positively to the question, "*can you tell us your overall impressions of using the system?*", and mentioned various benefits of the prototype. Moreover, eight (all participants except P7 and P10) showed their motivation for continued use toward the question: "*do you want to continue using the system?*" This result corroborated the support of H4, that is, the use of CatAlyst offers a favorable experience, and workers feel its usefulness. Moreover, four of them (P4, P5, P6, and P9) mentioned the novelty of the outputs as a key reason for their willingness to use the prototype.

> I've always been curious about what's generated. They were interesting. The excitement never stopped during the use and I'd not get bored. (P6)

These comments implied that the participants' favorable responses in Section 4 are not merely due to its freshness but thanks to the design of CatAlyst. Specifically, the unlimited variety in the outputs of the generative models plays a key role in keeping attracting workers' interest in the intervention, as we discussed in Section 3.2.1. The implications we obtained from the responses to the other questions are summarized below.

5.4.1 Effects on behavior. Seven participants (P2, P3, P4, P5, P6, P8, and P9) mentioned a significant change in their behavior owing to the system as an answer to the question: "*did you feel any change with and without the system?*"

Table 7: Backgrounds of the participants in Study 2.

	Gender	Writing habit	Main usage
P1	M	3 years	Novel
P2	F	8 years	Web article / Personal blog
P3	F	2 years	Diary
P4	M	0.5 years	Personal blog
P5	F	4 years	Web article / Personal blog
P6	F	rarely	University report
P7	M	10 years	Diary
P8	M	1 year	Fantasy / Lyric writing
P9	M	rarely	Private letter
P10	M	6 years	Novel

Table 8: Comparison of the interest retrieval time over the five days of the participants' use, which were not significantly different ($p = 0.791$).

day	1	2	3	4	5
# of detected interruptions	31	49	10	13	14
average retrieval time	95.4	117.1	121.1	105.2	93.4
standard deviation	81.9	91.4	102.7	95.4	89.6

> The notifications helped me cut out the time I spent touching my smartphone. I also sometimes started or proceeded to write sentences because I was curious to see what kind of sentences the AI would generate. I found this to be effective in the initial stages, when the writing was the heaviest. (P3)

The comments were parallel to those in the short study (Section 4.7) and demonstrated how CatAlyst affects workers and prevents their task procrastination. In particular, P3's comment implied a *catalyst*-like effect by lowering the hurdle for resuming the task, supporting our strategy.

5.4.2 Feelings about AI's accuracy. Next, the participants responded similarly to the question: "*what is your impression about the AI which generated content?*" More specifically, they pointed out the variance in the quality of the generated content.

> For my creative writing, the tool provided me with a good stimulus and motivation to continue writing because it allowed me to choose plots and words that I could not have reached on my own. However, I could not directly leverage it well for articulating my thought, like in a diary. (P8)

> I felt that AI has its strengths and weaknesses depending on the field of articles I write. As for abstract writing for hobbies, such as travelogues, novels, diaries, *etc.*, the AI helped me think about how to phrase words and what to do next, and it worked beyond my expectations. However, AI generation was not appropriate when it requires a great deal of checking, like using it for my job [web writing]. (P2)

> I was trying to write a full-length novel game scenario, and there was a constraint that each character setting was already fixed. While the outputs were interesting and I had fun seeing them, I needed to manually edit them, such as wording in generated utterances. (P10)

This result highlighted the difficulty of expecting a large generative model to perform well in all domains without any tuning, as we discussed in Section 2.1. However, it is noteworthy that the participants affirmatively perceived their experiences of using CatAlyst

owing to its design of not aiming to directly replace part of their tasks in the first place.

5.4.3 Role of CatAlyst. The participants' responses shed light on the three different roles of CatAlyst regarding the question: "*what was this system like for you?*"

CatAlyst as a reminder. Two participants (P2 and P9) appreciated the effective reminder function.

> I used the system for my job. Due to the working-from-home situation, I often get distracted. I appreciate the reminding function as well as its generated content, which lowers the hurdle for writing. (P2)

CatAlyst as an ideator. The benefit of obtaining ideas from CatAlyst was also emphasized by some participants (P1, P4, P8, and P10), who mostly performed creative writing, as follows:

> The suggested wording and story flow can be enough hints. For example, I got inspired to include a monologue of the protagonist or conversation to improve the tempo of the story. I used them to accelerate my writing. (P1)

This benefit is aligned with previous works supporting writing with generative models and confirms the validity of our expectation that CatAlyst can induce task resumption in workers by improving their ability to perform the task (Section 3.2.2).

CatAlyst as a peer. Other participants (P2, P3, P5, and P6) perceived CatAlyst as a peer or collaborator who motivated them:

> I found it like a co-worker working with me. I became curious about what they produce while I was taking a break. (P5)

> It is easier to expand my thinking by reviewing AI's outputs rather than contemplating from scratch by myself. In that sense, the AI was a peer to whom I can talk. (P3)

Their responses corresponded to our expectation that CatAlyst can gain workers' interest (Section 3.2.1) by stimulating their curiosity and can induce their task resumption by increasing their motivation (Section 3.2.2).

5.4.4 Room for further improvement of CatAlyst. The participants also mentioned their requests for the system when answering the question: "*what would you like to see with this system in the future?*" Some participants (P1, P4, and P7) mentioned the need to control notification timing as follows:

Ideally, I want to turn off the notification when I'm deeply thinking about the story flow, even if I'm not typing any words. (P1)

There is room for improving its accuracy because our detection approach is a practical but naïve method.

I hope the system cooperates with a smartwatch because I think it would be useful to receive notifications when I am walking or something like that, so I can think about it. (P4)

This comment implied a potential need for workers to actively set an intervention timing for CatAlyst as a reminder while performing different tasks (*e.g.*, walking). Moreover, there is an interesting conflict in the comments.

I could not wait for the AI's generation. It would have been even better if there had been a button for actively using it or more suggestions from the system. (P7)

I like its current design. If there had been an auto-generate button, it would have been a bad experience if the output had been inaccurate. In the current case, by thinking of it as an added feature of a reminder, my expectations were relaxed and I was glad to see it for the fun of it. (P6)

In Section 7.2, we further discuss this difference in the participants' needs and why we did not implement such a button for actively using generative models based on these comments.

Three participants (P1, P8, and P10) wanted to see multiple generations simultaneously.

I would appreciate it if it could suggest several patterns of sentences at the same time so that I can do a lot of thinking when I receive the notification. (P1)

This demand can be explained by prior research suggesting that presenting multiple perspectives, including AI's suggestions, can foster one's reflection by offering discussion grounds [4].

Additionally, two participants (P6 and P10) requested CatAlyst implemented in existing editors.

It's interesting, but I'd prefer to keep using my current editor [Google Docs] simply because it has more useful functions. If there is an API for integrating this system into my editor, I'd appreciate it. (P10)

In fact, CatAlyst can be easily integrated into online editors, as we will show a prototype in the form of a Chrome extension for supporting slide-editing on Google Slides in the next study.

6 STUDY 3: SLIDE-EDITING TASK

The short- and long-term studies in Sections 4 and 5 demonstrated the effectiveness of CatAlyst in writing tasks. We then conducted another study in which participants performed slide-editing tasks with the support of CatAlyst. This study aimed to examine the domain extensibility of our design, particularly when the large generative model used in CatAlyst was not tuned for this specific task. We again evaluated the hypotheses in Section 3.3 in a similar manner as in Section 4, a within-participant design involving the *proposed*, *control*, and *none* conditions.

6.1 Task

We prepared three tasks in which participants were asked to create a slide of at least 15 pages about an aging society, online abuse, or global warming. Here, we prepared a six-page draft slide for each topic and provided it to start the corresponding task. This aimed to balance the difficulty of the tasks because, without the draft, the quality of the slides and the time required to edit them would largely depend on each participant. It took roughly 30 min to complete a single task. We randomly assigned the three tasks to three conditions for each participant.

6.2 Implementation

In this study, we asked participants to use Google Slides[4] to edit the slides after installing a Chrome extension that we developed. The extension detects the moment when the participants become distracted according to their interaction log, in the same manner as in Section 4.3. In the *proposed* condition, the ID of the edited slide was sent to a remote server where its continuation was generated by transforming the slide in progress into a text input to the model. Here, we used the same GPT-3 model as the one used in Section 4.3 and we never tuned the model for slide-editing. In addition, we provided part of the model output to a diffusion model [58] to generate an image that complements the generated slide. The slide was then appended to the original slide via the Google Slides API[5], and the participants were prompted with a notification displaying the title of the generated slide. In the *control* condition, we used the same six encouraging messages as in Section 4.3 and showed a notification displaying one of them at the time of distraction.

6.3 Procedure

We followed the procedure of Study 1, involving twelve participants (eight males and four females; 20–59 years old[6]). Similarly, each participant completed four questionnaires: three at the end of each task and one at the end of all tasks. From their behavior during slide-editing and their responses to the questionnaires, we prepared the same measures as in Section 4.5, except for *subjective evaluation* and *progress after resumption*. For the subjective evaluation, we asked the raters to evaluate each slide with respect to consistency, visual appearance, and overall quality because readability is not applicable to this task. The progress after resumption is omitted because, at the interface of Google Slide, we could not obtain a precise edit history that allows us to quantify the progress either by the extension or by the API. Therefore, considering that H2 is supported in Section 4.6, we examined H1, H3, and H4 using the other measures.

6.4 Results

6.4.1 Interest retrieval time. Regarding H1, we obtained similar results as in Section 4.6; the interest retrieval time was significantly shorter ($t(165.05) = 3.16, p = 0.002$) for the *proposed* condition than for the *control* condition, as presented in Table 9.

---

[4]https://www.google.com/slides/about
[5]https://developers.google.com/slides/api
[6]We note that one participant preferred not to disclose their age.

Table 9: Comparison of the interest retrieval time between the *proposed* and *control* conditions, which were significantly different ($p = 0.002$).

	mean	SD
control	49.5	$\pm$ 81.8 (s)
proposed	19.2	$\pm$ 44.0 (s)

6.4.2 Ignorance rate. Moreover, as shown in Table 10, the ignorance rate was also significantly lower ($p < 0.001$). We also confirmed that, at the first intervention, the *control* condition was considerably effective compared to the *proposed* condition; however, repeated exposure to encouraging messages degraded their effectiveness.

6.4.3 Total editing time. For H3, a significant difference was observed in the time the participants spent between the three conditions, according to the Friedman test after Levene's test ($\chi^2(2) = 6.50, p = 0.039$), as shown in Table 11. We then conducted the post-hoc Nemenyi test [49] and found that the time spent in the *proposed* condition was significantly shorter than that in the *none* condition ($p = 0.038$).

6.4.4 Subjective quality. On the other hand, we did not observe a significant difference in the raters' evaluation of the quality of the slides the participants edited (Table 12), according to the Friedman test ($p = 0.311, 0.960,$ and 0.475 for consistency, visual appearance, and overall quality, respectively). Thus, it was implied that the *proposed* condition reduced the total time the participants spent while maintaining the quality of the edited slides. This suggests that CatAlyst can contribute to improving worker productivity, supporting H3.

6.4.5 Cognitive load. Based on the participants' responses to the questionnaire, we then examined their evaluations of the cognitive load. As illustrated in Figure 6, we found significant differences in cognitive load in terms of temporal demand ($F(2, 22) = 3.99, p = 0.033$), performance ($F(2, 22) = 3.79, p = 0.038$), effort ($F(2, 22) = 8.63, p = 0.001$), and frustration ($F(2, 22) = 7.80, p = 0.003$) according to ANOVA. The post-hoc test with Holm correction revealed that the scores of the effort and frustration factors in the *proposed* condition were significantly lower than those in the *control* ($t(11) = 3.60, p = 0.013$ and $t(11) = 3.04, p = 0.033$, respectively) and *none* ($t(11) = 2.79, p = 0.035$ and $t(11) = 2.9, p = 0.033$, respectively) conditions.

These results implied the advantage of our design. That is, the design of guiding the participants' interest using the generated continuation of the interrupted work, which is context-aware and variational, would ease their frustration. This would reduce the effort required to resume the task as well. Simultaneously, as mentioned in Section 4.7 and also later in Section 6.5, the generated continuation sometimes supported their progress by providing some ideas or reference information. This corroborated our observation that CatAlyst contributes to participants' perceived utility.

6.4.6 System usability. The participants' responses to items that measured system usability further supported this observation. As

Table 10: Comparison of the ignorance rate between the *proposed* and *control* conditions. While the rate was significantly lower for the proposed condition ($p < 0.001$), the difference was not observed when we focused on only the first intervention presented to each participant ($p = 0.712$).

	(A) **All**		(B) **First time**	
	ignored	worked	ignored	worked
control	85	104	7	11
proposed	19	70	4	11

presented in Table 13, the participants provided a higher SUS score for the *proposed* condition than for the *control* condition on average, although their difference was insignificant ($t(11) = 1.79, p = 0.101$). Furthermore, the three conditions exhibited a significant difference in the participants' rankings regarding usability, according to the Friedman test ($\chi^2(2) = 7.09, p = 0.029$). We found that the *proposed* condition was likely to be ranked significantly higher than the *control* condition ($p = 0.038$), as eight participants regarded the *proposed* condition as the best. This result supported H4, and together with the above results, we conclude that the effectiveness and extensibility of CatAlyst were confirmed.

6.5 User Comments

The comments of the participants highlighted the advantage of CatAlyst in drawing their interest and lowering the hurdle for resumption in the slide-editing task as well.

> I was surprised at the high quality of the automatically generated slides. I used the presented images as is and the provided text with a little editing.

> When I was notified with just a message, I was a little uncomfortable because I felt like being monitored. Regarding the notification from the AI, I was able to regard it as if the AI tried to give me an idea, so it was more satisfactory.

These comments are consistent with their evaluation of cognitive load.

However, some participants pointed out the limitation of the generated continuation, which might reflect the fact that the model used was not specifically tuned for slide-editing.

> As the quality of the AI's output was not always perfect, the process of deleting the automatically added slides or extracting useful parts from the slides and merging them into my own slides was somewhat laborious and demanding.

Still, the participant commented that "it was very helpful that the AI provided content ideas, which allowed me to focus consistently on the task without wondering." In other words, it was implied that large generative models that are not tuned for each individual task and might be imperfect can still be useful in helping workers avoid procrastination.
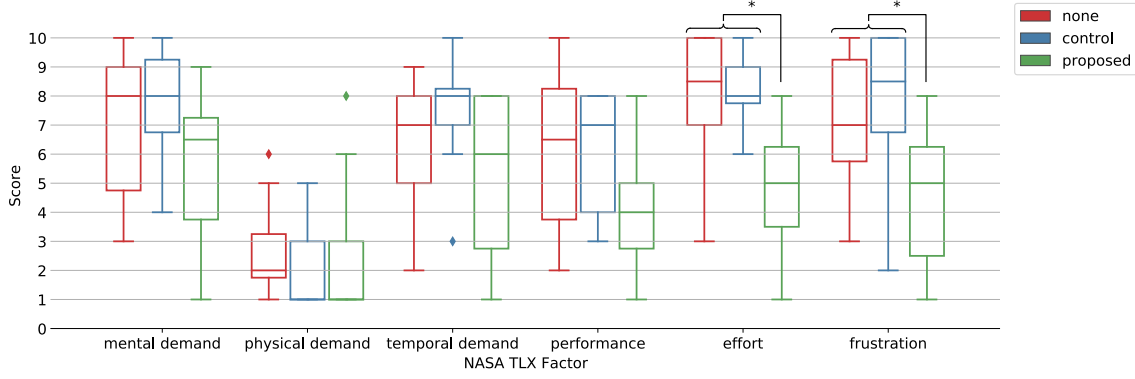
Figure 6: Participants' evaluations of their cognitive load in Study 3. The effort they made and the frustration they felt in the *proposed* condition was significantly lower than those of the *control* ($p = 0.013$ and 0.033, respectively) and *none* ($p = 0.035$ and 0.033, respectively) conditions.

Table 11: Comparison of the total editing time between the three conditions. The time spent by the participants in the *proposed* condition was significantly shorter than that in the *none* condition ($p = 0.038$).

	mean	SD
control	1776.1 ±	960.4 (s)
proposed	1356.5 ±	1048.7 (s)
none	2794.5 ±	2866.5 (s)

Table 12: Comparison of the subjective quality of the slides between the three conditions, which were not significantly different regarding consistency, visual appearance, and overall quality ($p = 0.311$, 0.960, and 0.475, respectively).

	consistency	visual appearance	overall quality
control	5.28 ± 1.09	5.00 ± 1.15	4.92 ± 1.16
proposed	5.22 ± 1.02	5.03 ± 1.18	5.08 ± 1.11
none	5.42 ± 0.87	4.89 ± 1.28	5.00 ± 1.20

7 FINDINGS AND DISCUSSION

Thus far, we have demonstrated the effectiveness and domain extensibility of CatAlyst by verifying our hypotheses through two different tasks (writing and slide-editing). Specifically, the results demonstrated CatAlyst can help workers avoid task procrastination with less cognitive load through interventions made by generative models without fine-tuning. At the same time, the series of studies informed us of further linkages between our study and prior work on human-AI collaborations. In particular, we found that the results highlighting the importance of reframing the role of generative models to induce workers' behavior change, rather than directly contributing to their tasks, corroborate the previous discussion. In this section, we discuss the findings of our study to clarify the linkages.

7.1 Exploiting Large Generative Models without Fine-tuning

Recent advances in large generative models are remarkable, such that they allow workers to delegate more work to AIs, raising expectations for achieving "the AI leads and the human assists" or even "full AI automation" [42]. Some of our results reconfirmed this capability; for instance, some participants appreciated the high-quality content generated by CatAlyst and actually incorporated it. Concurrently, the results highlighted the imperfections of the models and how they affect workers' experience, as the participants mentioned their insufficient quality in various writing domains in

Section 5.4.2. Even if the generation techniques continue to advance, it is not guaranteed that they work perfectly in diverse task domains. Thus, the wall of accuracy will remain in individual workers' tasks, which can be an inevitable issue when we aim for a higher level of task delegation; reported failure cases often include degraded trust and ignorance of the systems [32, 77, 78]. Tuning generative models to achieve higher accuracy is a theoretically valid solution; however, it is impractical because of its tremendous cost [3, 69], especially considering that there are many tasks and domains that each worker needs support for. Moreover, small datasets are often inadequate for tuning such large models [25, 52], which makes the solution intractable.

CatAlyst demonstrated an alternative way of contributing to the digital well-being of workers. In other words, our motivation for using generative models is different in the sense that it does not aim to increase task delegation but to influence workers' behavior. Our results exhibited that the ability of large generative models to produce context-aware content can be utilized for such interventions. More importantly, the results demonstrated its applicability in multiple domains without fine-tuning, suggesting a novel way of leveraging publicly available large models, in a bricolage manner, to support individual workers.

7.2 Passive Intervention vs. Active Use

An interesting conflict implied from the participants' responses in Section 5.4.4 is that some participants requested a button so that they could use AIs actively, while others mentioned concerns about

Table 13: Comparison of the participants' evaluations of the system usability between the *proposed* and *control* conditions, which were not significantly different (p = 0.101).

	mean SD
control	50.6 ± 19.4
proposed	68.8 ± 20.1

the idea. This discrepancy can be explained from the perspective of the relationship between expectations and trust toward AIs. As Yin *et al.* [81] suggested, workers' trust in the system would be affected by how much they expect in the system and how accurate the observed output is. In this sense, placing a button for actively using AIs can lead to workers' raised expectations, as they would rely on the feature to boost their work, and thus, their trust could be degraded if the output quality is insufficient. Therefore, in our design of CatAlyst, we did not include such buttons and limited the role of AIs to passive intervention for workers only when they were detected to be distracted. This idea aligns with the concept of *unremarkable AI* proposed by Yang *et al.* [80]. They designed a clinical decision support tool in which the existence of AI can only be noticed when it adds value to the decision while maintaining the existing decision-making routine. Similarly, in our case, we did not introduce generative models as *remarkable AI* within workers' task routines. We rather limited their presence outside of the original task (*i.e.*, when they are away from the task), which was supported by the participants. Simultaneously, their comments asking for buttons motivated us to conduct further research to examine the possibility of integrating the design of remarkable AI, while maintaining the current advantage of unremarkable AI. In addition, it is also possible to incorporate a design where generated continuations are continuously shown to workers and they can select whether to use them. We anticipate that such mixed-initiative interaction can lead to even better human-AI collaboration, as Lehmann and Buschek [36] discussed using the concept of autocomplete.

8   FUTURE WORK AND LIMITATIONS

Even though we verified the effectiveness of CatAlyst in two different tasks (writing and slide-editing), further research is demanded to test its domain extensibility. As we mentioned in Section 2.1, generative models have been proposed for other tasks, such as drawing. The advantage of CatAlyst, that is, its ability to leverage such models without tuning to individual tasks, motivated us to conduct studies in diverse situations involving workers with different proficiency levels and backgrounds.

We also want to explore how users of CatAlyst can incorporate the content of the generated continuations into their final outputs. This includes several aspects, from directly using sentences or images provided by the system to taking inspiration about possible ideas from them. Thus, the exploration would require an in-depth observation of the usage of CatAlyst alongside the current series of experiments, which will provide insightful discussions in combination with the findings of Lee *et al.* [35].

Moreover, while we used encouraging messages adopted from a public site blocker for the control condition, further comparison

with strategies of other prior work will provide more insights into the relationship between interventions and workers' behavior. For example, HabitLab [33] provides a rotation of diverse interventions to help users get back on task while maintaining their interest. It is, thus, a promising approach to use CatAlyst's intervention as one kind in the rotation to keep the novelty of the intervention.

One limitation is that our quantitative results are limited to laboratory studies in both tasks, similar to prior work [40]. This is due to the difficulty in recording the measures we used in a study conducted in the wild and making a fair comparison. For example, each worker would have tasks with different difficulties and priorities (*e.g.*, deadline, meeting, *etc.*). Therefore, in our in-the-wild study (Section 5), we focused on qualitatively confirming the usefulness of CatAlyst. It is desirable to conduct a longitudinal study measuring the digital well-being of workers as an outcome, as Howe and Menges did [30], and quantifying the effectiveness of CatAlyst for future work.

The mechanism of detecting workers' distractions can be further improved based on participants' feedback. The participants noticed that interventions sometimes occurred, even when they focused on the task. This is because our naïve method detected the status of workers based on the interaction log. Such methods are unable to consider moments when workers contemplate deeply (*e.g.*, thinking about a plot when writing a novel) without making any interactions on the web page. If we could implement software installed on their PCs, instead of a prototype web page, it would be possible to monitor more fine-grained activities, not just typing or not on a specific web page. Such information will help detect moments when workers are actually distracted with higher precision [72, 76], which would also contribute to quantifying the effectiveness of CatAlyst by contextualizing the moment in the longitudinal in-the-wild study mentioned above. Still, we want to emphasize that CatAlyst did not degrade workers' experiences compared to conventional encouraging messages through false-positive interventions. This is because CatAlyst can offer hints to foster their contemplation rather than alert them. Such false-positive resistance is an important aspect for machine-learning-based intervention to be successful [5] and exactly what CatAlyst aimed at, as mentioned in Section 3.1.2.

We also note that the use of large generative models poses a risk of bringing about some bias and ethical considerations [47]. For example, stereotypes and negative propagation of particular social groups have been observed in their outputs [9, 19, 64]. In CatAlyst, we cannot deny the possibility that such a bias is reflected in the final production of workers via the generated continuation. Toward deployment of CatAlyst, it is desirable to integrate measures to mitigate such bias [47] as well as provide a transparent explanation of the used models [48] and the clarification of the risk.

9   CONCLUSION

We presented a novel approach, CatAlyst, for supporting office workers in their task engagement through an intervention that leverages large generative models. Specifically, CatAlyst generates a continuation of interrupted work to prompt workers to resume the task by effectively attracting their interest and inducing behavior change for task resumption. The approach is domain-extensible, which means that we can develop systems on top of CatAlyst using

publicly available models without having to tune them to individual tasks. We conducted a series of studies to evaluate the effectiveness and usability of CatAlyst. The first study confirmed the effectiveness of CatAlyst in a writing task compared to a conventional encouraging approach. The second study further supported the favorable usability of CatAlyst throughout an in-the-wild study over a longer period by involving participants with different purposes for writing. The third study confirmed the effectiveness of CatAlyst in a slide-editing task, demonstrating its domain extensibility. Based on the results of the studies, we discussed the implications of our design, that is, leveraging large generative models to influence workers' behavior instead of delegating human tasks to AIs. Our findings and discussion pave the way for developing novel human-AI collaborations to enhance workers' digital well-being that can be harnessed in future human-AI symbiosis.

ACKNOWLEDGMENTS

REFERENCES

[1] Alexander Travis Adams, Jean Marcel dos Reis Costa, Malte F. Jung, and Tanzeem Choudhury. 2015. Mindless Computing: Designing Technologies to Subtly Influence Behavior. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, New York, NY, 719–730. https://doi.org/10.1145/2750858.2805843

[2] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 3:1–3:13. https://doi.org/10.1145/3290605.3300233

[3] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. In *Proceedings of the 2020 ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems.* arXiv, Ithaca, NY, 11 pages.

[4] Riku Arakawa and Hiromu Yakura. 2020. INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching. In *Proceeding of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 1–13. https://doi.org/10.1145/3313831.3376703

[5] Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 99:1–99:15. https://doi.org/10.1145/3411764.3445339

[6] Riku Arakawa and Hiromu Yakura. 2022. AI for human assessment: What do professional assessors need? *arXiv* abs/2204.08471 (2022). https://doi.org/10.48550/arXiv.2204.08471

[7] Joel Baer. 2021. Neural Filters: What Photoshop's Powerful New AI Tools can Do. https://blog.adobe.com/en/publish/2021/01/15/neural-filters-what-photoshop-s-powerful-new-ai-tools-can-do. Last accessed on September 1, 2022.

[8] Malikeh Beheshtifar, Hadis Hoseinifar, and Moghadam Nekoie Moghadam. 2011. Effect Procrastination on Work-Related Stress. *European Journal of Economics, Finance and Administrative Sciences* 38, 38 (2011), 59–64.

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, 610–623. https://doi.org/10.1145/3442188.3445922

[10] Daniel E Berlyne. 1960. *Conflict, Arousal, and Curiosity.* McGraw-Hill, New York, NY.

[11] Angela M. Zavaleta Bernuy, Ziwen Han, Hammad Shaikh, Qi Yin Zheng, Lisa-Angelique Lim, Anna N. Rafferty, Andrew Petersen, and Joseph Jay Williams. 2022. How can Email Interventions Increase Students' Completion of Online Homework? A Case Study Using A/B Comparisons. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference.* ACM, New York, NY, 107–118. https://doi.org/10.1145/3506860.3506874

[12] Angela M. Zavaleta Bernuy, Qi Yin Zheng, Hammad Shaikh, Andrew Petersen, and Joseph Jay Williams. 2021. Investigating the Impact of Online Homework Reminders Using Randomized A/B Comparisons. In *Proceedings of the he 52nd*

[13] Petter Bae Brandtzæg and Asbjørn Følstad. 2017. Why People Use Chatbots. In *Proceedings of the 4th International Conference on Internet Science*, Vol. 10673. Springer, Cham, Switzerland, 377–392. https://doi.org/10.1007/978-3-319-70284-1_30

[14] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). CRC Press, London, UK, 207–212.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 2020 Annual Conference on Neural Information Processing Systems.* Curran Associates, Red Hook, NY, 1877–1901.

[16] James C. Byers, Alvah C. Bittner, and Susan G. Hill. 1989. Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparisons Necessary?. In *Proceedings of the 1989 Annual International Industrial Ergonomics and Safety Conference.* Taylor & Francis, Philadelphia, PA, 481–485.

[17] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, New York, NY, 2287–2295. https://doi.org/10.1145/3292500.3330723

[18] Hyunsung Cho, Daeun Choi, Donghwi Kim, Wan Ju Kang, Eun Kyoung Choe, and Sung-Ju Lee. 2021. Reflect, not Regret: Understanding Regretful Smartphone Use with App Feature-Level Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 456:1–456:36. https://doi.org/10.1145/3479600

[19] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *arXiv* 2202.04053 (2022), 20 pages.

[20] Francesco Cirillo. 2018. *The Pomodoro Technique: The Life-Changing Time-Management System.* Random House, New York, NY.

[21] James Clear. 2015. The Chemistry of Building Better Habits. https://jamesclear.com/chemistry-habits. Last accessed on September 2022.

[22] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* ACL, Stroudsburg, PA, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[23] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. *arXiv* 2208.09323 (2022), 13 pages. https://doi.org/10.48550/arXiv.2208.09323

[24] Roelof Anne Jelle de Vries, Cristina Zaga, Franciszka Bayer, Constance H. C. Drossaert, Khiet P. Truong, and Vanessa Evers. 2017. Experts Get Me Started, Peers Keep Me Going: Comparing Crowd- Versus Expert-Designed Motivational Text Messages for Exercise Behavior Change. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare.* ACM, New York, NY, 155–162. https://doi.org/10.1145/3154862.3154875

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* ACL, Stroudsburg, PA, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[26] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.1037/xge0000033

[27] B. J. Fogg. 2009. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology.* ACM, New York, NY, 1–7. https://doi.org/10.1145/1541948.1541999

[28] Michael D. Frakes and Melissa F. Wasserman. 2016. *Procrastination in the Workplace: Evidence from the US Patent Office.* Technical Report. National Bureau of Economic Research.

[29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183. https://doi.org/10.1016/s0166-4115(08)62386-9

[30] Lauren C. Howe and Jochen I. Menges. 2021. Remote Work Mindsets Predict Emotions and Productivity in Home Office: A Longitudinal Study of Knowledge Workers During the COVID-19 Pandemic. *Human–Computer Interaction* in press (2021), 1–27. https://doi.org/10.1080/07370024.2021.1987238

[31] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm

ACM Technical Symposium on Computer Science Education. ACM, New York, NY, 921–927. https://doi.org/10.1145/3408877.3432427

Aversion. In *Proceedings of the 28th European Conference on Information Systems*. AIS, Atlanta, GA, 1–16.

[32] Jingoog Kim, Mary Lou Maher, and Safat Siddiqui. 2021. Studying the Impact of AI-based Inspiration on Human Ideation in a Co-Creative Design System. In *Proceedings of the 2nd ACM IUI Workshops on Human-AI Co-Creation with Generative Models*. CEUR-WS.org, Aachen, Germany, 16 pages.

[33] Geza Kovacs, Zhengxuan Wu, and Michael S. Bernstein. 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 95:1–95:25. https://doi.org/10.1145/3274364

[34] Clarry H. Lay. 1986. At Last, My Research Article on Procrastination. *Journal of Research in Personality* 20, 4 (1986), 474–495. https://doi.org/10.1016/0092-6566(86)90127-3

[35] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 388:1–388:19. https://doi.org/10.1145/3491102.3502030

[36] Florian Lehmann and Daniel Buschek. 2021. Examining Autocompletion as a Basic Concept for Interaction with Generative AI. *i-com* 19, 3 (2021), 251–264. https://doi.org/10.1515/icom-2020-0025

[37] Da-Wei Li, Danqing Huang, Tingting Ma, and Chin-Yew Lin. 2021. Towards Topic-Aware Slide Generation For Academic Papers With Unsupervised Mutual Learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, 13243–13251.

[38] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–14. https://doi.org/10.1145/3313831.3376258

[39] Erin Chao Ling, Iis Tussyadiah, Aarni Tuomi, Jason Stienmetz, and Athina Ioannou. 2021. Factors Influencing Users' Adoption and Use of Conversational Agents: A Systematic Review. *Psychology & Marketing* 38, 7 (2021), 1031–1051. https://doi.org/10.1002/mar.21491

[40] Yikun Liu, Yuan Jia, Wei Pan, and Mark S. Pfaff. 2014. Supporting Task Resumption Using Visual Feedback. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, 767–777. https://doi.org/10.1145/2531602.2531710

[41] 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*. ACM, New York, NY, 1–13. https://doi.org/10.1145/3313831.3376739

[42] Brian Lubars and Chenhao Tan. 2019. Ask Not What AI Can Do, but What AI Should Do: Towards a Framework of Task Delegability. In *Proceedings of the 2019 Annual Conference on Neural Information Processing Systems*. Curran Associates, Red Hook, NY, 57–67.

[43] Norman H. Mackworth. 1948. The Breakdown of Vigilance During Prolonged Visual Search. *Quarterly Journal of Experimental Psychology* 1, 1 (1948), 6–21. https://doi.org/10.1080/17470214808416738

[44] Chandra Shekhar Maddila, Sai Surya Upadrasta, Chetan Bansal, Nachiappan Nagappan, Georgios Gousios, and Arie van Deursen. 2022. Nudge: Accelerating Overdue Pull Requests Towards Completion. *ACM Transactions on Software Engineering and Methodology* 2011.12468 (2022), 30 pages. https://doi.org/10.1145/3544791

[45] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 107–110. https://doi.org/10.1145/1357054.1357072

[46] Paul M Mastrangelo, Wendi Everton, and Jeffery A Jolton. 2006. Personal Use of Work Computers: Distraction Versus Destruction. *CyberPsychology & Behavior* 9, 6 (2006), 730–741.

[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 115:1–115:35. https://doi.org/10.1145/3457607

[48] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 220–229. https://doi.org/10.1145/3287560.3287596

[49] Peter Nemenyi. 1963. *Distribution-Free Multiple Comparisons*. Ph.D. Dissertation. Princeton University.

[50] Raja Parasuraman. 1979. Memory Load and Event Rate Control Sensitivity Decrements in Sustained Attention. *Science* 205, 4409 (1979), 924–927.

[51] Evandro Morais Peixoto, Ana Celi Pallini, Robert J Vallerand, Sonia Rahimi, and Marcus Vinicius Silva. 2021. The Role of Passion for Studies on Academic Procrastination and Mental Health During the COVID-19 Pandemic. *Social Psychology of education* 24, 3 (2021), 877–893. https://doi.org/10.1007/s11218-021-09636-9

[52] Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *CoRR* 1811.01088 (2018), 12 pages. https://doi.org/10.48550/arXiv.1811.01088

[53] Timothy A Pychyl. 2013. *Solving the Procrastination Puzzle: A Concise Guide to Strategies for Change*. TarcherPerigee, New York, NY.

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 140:1–140:67.

[55] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 314:1–314:7. https://doi.org/10.1145/3411763.3451760

[56] Joaquin Alfredo Rodriguez, Gabriele Piccoli, and Marcin Bartosiak. 2019. Nudging the Classroom: Designing a Socio-Technical Artifact to Reduce Academic Procrastination. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. ScholarSpace, Honolulu, HI, 1–10.

[57] Melissa Roemmele and Andrew S. Gordon. 2015. Creative Help: A Story Writing Assistant. In *Proceedings of the 8th International Conference on Interactive Digital Storytelling*, Vol. 9445. Springer, Cham, Switzerland, 81–92. https://doi.org/10.1007/978-3-319-27036-4_8

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation (CVF) / IEEE, New York, NY, 10684–10695.

[59] C. Douglas Saddler and Laurie A. Sacks. 1993. Multidimensional Perfectionism and Academic Procrastination: Relationships with Depression in University Student. *Psychological Reports* 73, 3 (1993), 863–871. Issue 1. https://doi.org/10.1177/00332941930733pt123

[60] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. PEER: A Collaborative Language Model. *arXiv* 2208.11663 (2022), 22 pages. https://doi.org/10.48550/arXiv.2208.11663

[61] Henri C. Schouwenburg. 1992. Procrastinators and Fear of Failure: An Exploration of Reasons for Procrastination. *European Journal of Personality* 6, 3 (1992), 225–236. https://doi.org/10.1002/per.2410060305

[62] Athar Sefid, Jian Wu, Prasenjit Mitra, and C. Lee Giles. 2019. Automatic Slide Generation for Scientific Papers. In *Proceedings of the 3rd International Workshop on Capturing Scientific Knowledge*, Vol. 2526. CEUR-WS.org, Aachen, Germany, 11–16.

[63] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. 2021. SAGA: Collaborative Storytelling with GPT-3. In *Companion Publication of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, 163–166. https://doi.org/10.1145/3462204.3481771

[64] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, Stroudsburg, PA, 3405–3410. https://doi.org/10.18653/v1/D19-1339

[65] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction in press* (2022), 53 pages. https://doi.org/10.1145/3511599

[66] Konrad Sowa, Aleksandra Przegalinska, and Leon Ciechanowski. 2021. Cobots in Knowledge Work. *Journal of Business Research* 125 (2021), 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

[67] Aarohi Srivastava et al. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv* abs/2206.04615 (2022). https://doi.org/10.48550/arXiv.2206.04615

[68] Jan Willem Streefkerk, Myra P. van Esch-Bussemakers, and Mark A. Neerincx. 2008. Field Evaluation of a Mobile Location-Based Notification System for Police Officers. In *Proceedings of the 10th Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, New York, NY, 101–108. https://doi.org/10.1145/1409240.1409252

[69] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, 13693–13696.

[70] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J. Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI During Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 582:1–582:11. https://doi.org/10.1145/3411764.3445219

[71] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven, CT.

[72] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting Student Engagement in Classrooms Using Facial Behavioral Cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*. ACM, New York, NY, 33–40.

[73] Bin Wang, Yukun Liu, Jing Qian, and Sharon K. Parker. 2020. Achieving Effective Remote Working During the COVID-19 Pandemic: A Work Design Perspective. *Applied Psychology* 70, 1 (2020), 16–59. https://doi.org/10.1111/apps.12290

[74] Chatchai Wangwiwattana, Sunjoli Aggarwal, and Eric C. Larson. 2021. Writers Gonna Wait: The Effectiveness of Notifications to Initiate Aversive Action in Writing Procrastination. *arXiv* 2101.10191 (2021), 13 pages. https://doi.org/10.48550/arXiv.2101.10191

[75] Christopher D. Wickens. 2021. Attention: Theory, Principles, Models and Applications. *International Journal of Human-Computer Interaction* 37, 5 (2021), 403–417. https://doi.org/10.1080/10447318.2021.1874741

[76] Hiromu Yakura, Tomoyasu Nakano, and Masataka Goto. 2022. An Automated System Recommending Background Music to Listen to While Working. *User Modeling and User-Adapted Interactions* 32, 3 (2022), 355–388.

[77] Chuan Yan, John Joon Young Chung, Yoon Kiheon, Yotam I. Gingold, Eytan Adar, and Sungsoo Ray Hong. 2022. FlatMagic: Improving Flat Colorization through AI-driven Design for Digital Comic Professionals. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 380:1–380:17. https://doi.org/10.1145/3491102.3502075

[78] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing. In *Proceedings of the 3rd ACM IUI Workshops on Human-AI Co-Creation with Generative Models*. CEUR-WS.org, Aachen, Germany, 56–65.

[79] Qian Yang, Aaron Steinfeld, Carolyn Penstein Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. https://doi.org/10.1145/3313831.3376301

[80] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 238. https://doi.org/10.1145/3290605.3300468

[81] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 279. https://doi.org/10.1145/3290605.3300509

[82] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 841–852. https://doi.org/10.1145/3490099.3511105

[83] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child-AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 311:1–311:15. https://doi.org/10.1145/3491102.3501914

[84] Chao Zhang, Zili Zhou, Jiayi Wu, Yajing Hu, Yaping Shao, Jianhui Liu, Yuqi Hu, Fangtian Ying, and Cheng Yao. 2021. Bio Sketchbook: An AI-assisted Sketching Partner for Children's Biodiversity Observational Learning. In *Proceedings of the 20th ACM Interaction Conference on Design and Children, Athens*. ACM, New York, NY, 466–470. https://doi.org/10.1145/3459990.3465197

[85] Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. 2022. Telling Stories from Computational Notebooks: AI-Assisted Presentation Slides Creation for Presenting Data Science Work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 53:1–53:20. https://doi.org/10.1145/3491102.3517615

[86] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity Assessment of Neural Code Completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. ACM, New York, NY, 21–29. https://doi.org/10.1145/3520312.3534864