

## AI-Assisted Human Labeling: Batching for Efficiency without Overreliance

ZAHRA ASHKTORAB, IBM Research, USA  
 MICHAEL DESMOND, IBM Research, USA  
 JOSH ANDRES, IBM Research, Australia  
 MICHAEL MULLER, IBM Research, USA  
 NARENDRA NATH JOSHI, IBM Research, USA  
 MICHELLE BRACHMAN, IBM Research, USA  
 AABHAS SHARMA, IBM Research, USA  
 KRISTINA BRIMIJOIN, IBM Research, USA  
 QIAN PAN, IBM Research, USA  
 CHRISTINE T. WOLF, Independent Researcher, USA  
 EVELYN DUESTERWALD, IBM Research, USA  
 CASEY DUGAN, IBM Research, USA  
 WERNER GEYER, IBM Research, USA  
 DARREL REIMER, IBM Research, USA

Human labeling of training data is often a time-consuming, expensive part of machine learning. In this paper, we study "batch labeling", an AI-assisted UX paradigm, that aids data labelers by allowing a single labeling action to apply to multiple records. We ran a large scale study on Mechanical Turk with 156 participants to investigate labeler-AI-batching system interaction. We investigate the efficacy of the system when compared to a single-item labeling interface (i.e., labeling one record at-a-time), and evaluate the impact of batch labeling on accuracy and time. We further investigate the impact of AI algorithm quality and its effects on the labelers' overreliance, as well as potential mechanisms for mitigating it. Our work offers implications for the design of batch labeling systems and for work practices focusing on labeler-AI-batching system interaction.

---

Authors' addresses: Zahra Ashktorab, Zahra.Ashktorab1@ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Michael Desmond, mdesmond@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Josh Andres, josh.andres@au1.ibm.com, IBM Research, W Tower, 60 City Rd Southgate, Melbourne, Vic, 3006, Australia; Michael Muller, michael\_muller@us.ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Narendra Nath Joshi, Narendra.Nath.Joshi@ibm.com, IBM Research, 650 Harry Road Almaden Research Center, San Jose, CA, 95120, USA; Michelle Brachman, michelle.brachman@ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Aabhas Sharma, Aabhas.Sharma@ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Kristina Brimijoin, kbrimij@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Qian Pan, Qian.Pan@ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Christine T. Wolf, chris.wolf@gmail.com, Independent Researcher, San Jose, CA, , USA; Evelyn Duesterwald, duester@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Casey Dugan, cadugan@us.ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Werner Geyer, werner.geyer@us.ibm.com, IBM Research, 75 Binney Street, Cambridge, MA, 02142, USA; Darrel Reimer, dreimer@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2021 Association for Computing Machinery.

2573-0142/2021/4-ART89 \$15.00

<https://doi.org/10.1145/3449163>

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **Natural language interfaces**; *Interactive systems and tools*; *Empirical studies in interaction design*; User studies; User interface design.

Additional Key Words and Phrases: AI, Agents, Collaboration, Data labeling

### ACM Reference Format:

Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T. Wolf, Evelyn Duesterwald, Casey Dugan, Werner Geyer, and Darrel Reimer. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 89 (April 2021), 27 pages. <https://doi.org/10.1145/3449163>

## 1 INTRODUCTION

Labeling training data has long imposed a bottleneck on many machine learning tasks. It is a required step to train supervised machine learning algorithms [7]. Labeling is both a time-intensive and expensive process in which crowdworkers and/or subject matter experts must use their cognitive abilities to categorize and label vast datasets - often a repetitive process. Crowdsourcing platforms offer an abundance of relatively cheap labor and are widespread for labeling tasks [4, 29, 37]. Subject Matter Expert labeling frameworks also exist [54], but there the labor force is small and likely extremely expensive. Crowd workers as well as Subject Matter Experts could potentially benefit from tools to make their tasks more efficient and less laborious. Many researchers have investigated how to design labeling tasks to help people label more quickly and more efficiently. In this paper, we study “batch labeling,” where a single labeling action can be applied to multiple records. We investigate how AI-driven batching of items impacts labeler performance. To do this, we need to understand how individuals *perceive* the quality of the batches that are produced by an Artificial Intelligence (AI) system, and also measure the objective coherence of AI-generated batches. We also need to be wary of the AI system working “too hard” on behalf of the labelers, leading to the risk that they may *overrely* [47] on the AI assistance, and fail to exercise human discernment when assigning labels to items in a batch.

In this paper we investigate a batching tool that leverages AI to make batched recommendations for those who are doing the labeling work. We conducted four studies to explore the following six research questions:

- RQ1:** Does batching impact the outcome (accuracy score and time spent on task of the labeling task)?
- RQ2:** How does task complexity (vocabulary size) impact the outcome of data labeling?
- RQ3:** Does telling users that the batches are generated by an AI system impact the outcome (accuracy score, time spent on task agreement with batches, and overreliance)?
- RQ4:** How does the quality of the AI generated batches impact the outcome (accuracy score, time spent on task agreement with batches, and overreliance) of the labeling task?
- RQ5:** Can we mitigate overreliance on AI-assisted batching by asking users to rate the batches suggested by the AI system?
- RQ6:** Can we mitigate overreliance on AI-assisted batching by signaling to users that their responses will help improve the AI’s batch recommendations in the future?

We pursued these questions through four experiments.

- In Study 1 (**RQ1, RQ2**), we investigate the effects task complexity and batching have on data labeling performance (accuracy and time taken to complete task)
- In Study 2 (**RQ3**), we investigate the impact of the feedback given to the labeler about AI’s involvement in the process on users’ data labeling performance. We also examine labeler agreement with batches and overreliance on its recommendations.

- In Study 3, (RQ4), we investigate the impact of the quality of the AI generated batches on user performance.
- In Study 4 (RQ5, RQ6), we investigate overreliance mitigation strategies and their impact on data labeling performance and agreement/overreliance.

The contributions of this paper are the following:

- We demonstrate batching is more effective for complex tasks than simple tasks when looking at total time spent on task
- We demonstrate that AI quality impacts both the degree to which labelers agree with batching recommendations and their overreliance on recommendations.
- We provide design implications for building and designing effective AI-driven batching systems for labeling. These implications include:
  - i. Baseline settings for “Select-All” buttons on batching systems
  - ii. Dynamically allowing users to choose batch sizes or generating batch sizes based on item similarity
  - iii. Providing thoughtful feedback to promote mindfulness during the labeling task
  - iv. Providing label suggestions in addition to batching

## 2 RELATED WORK

### 2.1 Automated and Labeling Assistance Approaches

Labeling training data is an expensive and time-consuming process. Prior work has explored creative approaches to gain time efficiencies. For example, Ratner et al. [51] investigated the idea of *data programming* by allowing users to write labeling functions that express arbitrary heuristics which can reveal accuracies and correlations as a starting training data set. While this approach improves time efficiency, it requires labelers to learn how to write labeling functions. In the biomedical domain, Fauqueur et al. [17] explored a method to generate weakly labeled data in an automated manner for the construction of knowledge bases to derive biomedical relationships. This approach aims to combine automatic data generation with domain expert feedback. Vajda et al. [65] explored a semi-automatic labeling technique for large handwritten character collections by taking an unsupervised clustering and minimal expert knowledge approach. Their work suggests that by using their system, the involvement of human labor can be lessened to obtain cost gains; however, labeling data remains a lengthy process. Demirkus et al. [13] developed a semi-automatic framework for labeling temporal head poses in real-world video sequences, a task that is both time-intensive and expensive. Their framework automates the labeling process for video-labeling and achieves a 96.8% accuracy over manual labeling video frames (30%). Getting quality work from crowdworkers is a challenge, as demonstrated by prior work that investigated ways to get quality data from crowdworkers through aptitude tests and training in qualitative coding, showing that in-person training can improve quality of data collected on crowdsourcing platforms [42].

These works show the range of ideas being explored to assist labelers with labeling data more accurately and faster through automated techniques and non-automated techniques. However, the successful work with images and videos does not necessarily transfer into the domain of text. Images and videos have a predictable “surface” structure, whereas text often deals with *latent* matters of semantics. AI-driven batching recommendation interfaces have not yet been explored to improve the labeling process.

### 2.2 Human-AI Interaction

The growing prevalence of contemporary AI in a wide array of everyday systems has called our attention to the nature and experience of interactions between humans and intelligent machines [14].

Early AI systems (based largely on planning-based techniques) were critiqued for their inability to adapt in face of emergent, situated action [62]. Our technological encounters gradually became “less a matter of pushing buttons and pulling levers” and more about situated interaction, dialogue, and conversation between people and machines [26, 63]. Later work, most notably the “mixed initiative” approach, pointed to a number of design principles to foster more intuitive and adaptive interactions between humans and intelligent systems; a central theme in these design principles is the need to grapple with uncertainty and that humans’ goals can evolve through action [26]. Contemporary AI, driven by Big Data and statistical machine learning, raises important questions for social computing researchers as the promise of machines as collaborative partners in a number of everyday and domain settings becomes more and more a reality. A number of CSCW papers in recent years have addressed questions of human-AI collaboration, for example in medical decision-making [8], data science work [66], or IT infrastructure design practices [72].

Recent work in the HCI and CSCW community investigates how users try to understand and form impressions of AI systems and outcomes in human-AI collaborative settings [3, 67]. While a recent survey found AI-decision-making systems to be unfavorable with people [2], other studies have found that in many domains individuals believe in the inevitability of human-AI collaboration and were optimistic about the potential in data science [66], music [11], pedagogical tools for health literacy [61] and other forms of tutoring [1]. While computer-mediated communication research has extensively studied how users form impressions of other people they are communicating and interacting with through computer mediated channels [27], there have been few studies investigating user performance and overreliance on AI technologies in AI-driven labeling systems.

**2.2.1 Labeling as Collaborative Work.** Data science, and particularly machine learning, are increasingly carried out by teams [35, 43, 48]. The 2017 Kaggle survey identified 15 distinct roles in data science teams, such as data scientist (15%), software developer (11%) data analyst (7%), scientist/researcher (6%), business analyst (5%), and additional roles with lower frequencies of occurrence [28]. The complex nature of data science projects makes it unlikely that a single person will have sufficient knowledge or skills to create or implement all aspects of a project [30, 40]. Collaboration practices among diversely-skilled members of data science teams may be complex, requiring distinct tools, and involving nuanced communication patterns [76].

The work of data labelers (or annotators) is part of this complexity, and the tools used by labelers constitute a family of semi-specialized applications (for a review of applications, see Section 2.4, below). In many domains, a data scientist or a domain expert specifies a vocabulary of labels, and the task of applying those labels to instances (data records) falls to other members of the data science team [19, 68, 69]. After the labels have been added to data records, then the dataset itself is passed along to other roles in the complex hierarchy of data science teams [28] and their work-practices [76]. Thus, labels are collaborative objects, and labelers make their contributions within the collaborative fabric of data science teams.

### 2.3 Overreliance on Automation and Artificial Intelligence

Automation bias is defined as the “tendency to over-rely” on automated systems [21]. This issue has been investigated in a number of critical industries, such as medicine [21] and aviation [52]. In a situation of overreliance, humans uncritically follow recommendations from an automated system, without using their own perceptual capacities and judgment. Overreliance can be harmful because it degrades human-in-the-loop processes - where the human plays an important role in providing feedback, shaping machine action, and acting as a monitor or quality assurance to ensure “all goes well” [58].

More and more processes are becoming automated due to advances in AI, for example, from label recommendation systems for data labelers when labeling data [65], to an automatic aeroplane control for pilots when flying planes [6]. One challenge emerging as automation is introduced within the workflow of humans and machines working together is that of the human over-relying on the system capabilities [23]. Tragic examples are known from airplane crews being unable to coordinate with the automated system in how to respond to a situation [20, 60]. Human overreliance on the system is problematic and can contribute to users losing critical engagement with the process [23]. Overreliance on automation has historically been studied by measuring reactions by trained pilots on automated aircraft control tasks that would occasionally fail [52].

Reliance and Trust are closely related constructs. Many instruments that measure trust ask how likely an individual is to follow the recommendations given by that system [12, 31, 64]. In a recent study, user trust was measured based on how likely an individual was to follow the recommendations given by an AI [75]. Trust is an important aspect of human-AI interaction. Individuals may decide not to use a reliable, well-designed system if they do not trust it. Conversely, they might use an unreliable system if they do trust it. Both of these phenomena have been observed in prior work [36]. Overreliance and complacency of automated reliable systems has been documented in prior work [47]. Complacency on automated systems has been described as the consistency of the reliability of the automated task in a multi-task environment [46]. Prior work identifies conditions that lead to complacency or the premature cognitive commitment to an automated device as: routine, repetition and extremes of workload [32] - all conditions that occur in labeling tasks. To our knowledge, overreliance has not yet been studied in AI-assisted labeling systems.

Prior work has documented different ways in which complacency and overreliance can be addressed. Miller et al. have suggested that increasing a user's awareness of the situation and system performance and giving that user more control of how much automation to use, may ameliorate complacency [41]. Historically, airplane pilot tests have attempted to address overreliance on automating pilot tests [52], since in the worst case scenario of an example of overreliance, a pilot can crash an airplane by overrelying on the airplane automation, as seen in a crash in Columbus, Ohio in 1994 [47]. More recent work has attempted to make decision making machine learning processes more transparent and interpretable to help users come to an informed decision about how much they want to rely on the AI decision making system [22]. In this paper, we explore mitigations in our AI-driven batching system to study the mitigation of labeler overreliance on AI-recommended batches.

## 2.4 Machine-Teaching

In a simple sense, any labeling or annotating task involves a human teaching a machine-learning model about the ground truth of the data [24, 49]. It might be argued that each labeling task is in some way "unique." However, a growing family of labeling and annotation tools have been built and used across *diverse* media, including images, audio passage, chatbot dialogs, medical records, and other domains. The BRAT tool [59] provides a good example of generalizability. It was originally designed for use in linguistics, but has also been applied in studies of healthcare [73] and rhetoric [5]. Similarly, CrowdFlower has been used as the UI component in studies of linguistics [74], social media [18], and audio recordings [9]. However, all of these tools depend primarily on human abilities. AI tools show promise to extend and augment humans' labeling capabilities.

As the relationship between AI and humans becomes more collaborative [58], researchers continue to investigate different ways to make these relationships yield a better user experience. Interactive machine teaching is a growing area that sees the ML model training process through a human-centered lens [15]. The interactive machine teaching paradigm is seen as making ML building practices more accessible to novices and non-ML experts by framing the training process

in terms of a collaborative relationship between humans and ML systems [15, 50]. Hong et al [25] used a machine teaching exercise (e.g., mTurk users uploaded photos to teach the model various categories) to understand how human teachers understand and reflect on their experiences with ML. They found that participants understood the need for diversity in training samples, but also struggled with understanding notions of consistency in training samples, as well as misconceptions about the model's reasoning capabilities. The mitigation deployed later in this study leverages this prior machine teaching work by framing the labeling process as a collaborative one between the user and the AI batching system.

The screenshot shows a task interface for 'Item # 1' with '0/240 labeled'. The 'Question' section contains the instruction: 'Read the example text and then select the label that best describes it best below'. The example text is: 'Where can I find my bank routing numbers for money transfer?'. The 'Labels' section has a dropdown menu with 'Banking' and 'Telecommunication and Phone Services' options. A 'Submit' button is located below the labels.

Fig. 1. The simple task x no batching condition seen by participants in Group 0 in Study 1 (see Table 1). A) where batching occurs/does not occur and B) where labels are listed reflecting task complexity.

The screenshot shows a task interface for '0/60 labeled'. The 'Current Items' section shows a list of items: 'Item 0', 'Item 1', 'Item 2', and 'Item 3'. The 'Question' section contains the instruction: 'Read the example text and then select the label that best describes it best below'. The example text is: 'Where can I find my bank routing numbers for money transfer?'. The 'Labels' section has a dropdown menu with a list of labels: 'Activate Bank Card', 'Activate Phone/Device', 'Activate Roaming for Phone Device', 'Cancel Bank Card', 'Changing/Keeping Phone Number', 'Fee Inquiry from Bank', 'Replace Bank Card', 'Report Missing Bank Card', 'Return a Device/Phone', 'Swap a Device/Phone', 'Unlocking a Device/Phone', and 'View Bank Routing Number'. A 'Submit' button is located below the labels.

Fig. 2. The complex task x batching condition seen by participants in Group 3 in Study 1 and 2 (see Table 1). A) batching occurs/does not occur and B) where labels are listed reflecting task complexity.

## 2.5 Active learning

Using unsupervised machine learning to batch data for human labeling efficiency shares some general characteristics with active learning [57]. Active learning is a human-in-the-loop labeling paradigm wherein the 'learner' (generally a supervised model) is used to select which examples to label from a pool of unlabeled data. A common implementation is to use the learner's 'uncertainty' as a data selection heuristic. Both active learning and batch labeling utilize a machine learning algorithm to organize and optimize the human labeling task. In the case of batch labeling, the algorithm is unsupervised, and the goal is to reduce redundant labeling effort by grouping data for presentation to the labeler. Active learning combines the output of a supervised model with a selection heuristic to minimize the overall amount of data that is necessary to label. The way in

which both paradigms treat data is subtle but important. However, active learning does not affect the labeling experience as deeply as batching. While it has some properties that could affect user experience (such as difficult of labeling), its utility focuses on the order and quantity of data that needs to be labeled.

In this paper, we focus on batching as an interaction design and the impact it has on the outcome of labeling. Future investigation might concentrate on integrating active learning into the batch labeling process as it would likely improve labeling performance at the cost of a more complex implementation. Ordering of batches using an active learning selection strategy may lead to a more optimal labeling process, whereby the most uncertain batches are labeled first, and the overall number of batches is reduced. A more advanced integration might dynamically use heuristic data in the construction of batches themselves.

### 3 SYSTEM DESCRIPTION

#### 3.1 Overview of System

The studies presented in this paper were conducted on a data labeling platform we built (shown in Figures 1 and 2). Our motivation for building a new labeling platform, as opposed to using one of the existing labeling/crowd-working tools in the market, is two-fold. First, we envision our platform to primarily cater to Subject Matter Experts (SMEs) who bring their domain of expertise to drive the data labeling process, ensuring the high quality and satisfactory justification of the labeled data. Second, designing our own data labeling platform allows us to study and fine-tune the AI system's labeling assistance during the user experience. The AI assistance feature we focus on in this study is the paradigm of batch labeling which is discussed in detail in Section 3.2

Our system is set up to give users the ability to create custom data labeling projects with granular control over various configurations of the labeling task. These configurations include, among others, project metadata (name, description, etc.), whether AI-driven batches are enabled, label taxonomy size, whether single or multiple labels can be applied to a data item, and the labelers who will be invited to perform labeling tasks on this project. As part of project creation, project owners upload and preview the dataset to be labeled, configure the labeling task prompt and input the labeling taxonomy. Project owners have the administrator privileges to modify and delete projects as needed. Once a project is configured and launched, project owners and labelers can begin labeling data items via the interface pictured in Figure 1 and Figure 2, depending on the configuration. The data items are presented on the left side of the screen while the labeling question and label options are posed on the right side of the screen. Depending on project configuration, labelers can submit one label or multiple labels. If the project is configured to enable AI-driven batches, the data items are presented in batches as explained in Section 3.2

#### 3.2 Batch Labeling: Baseline AI

Our data labeling platform includes "batch labeling", an AI-assisted UX paradigm, that aids data labelers by allowing a single labeling action to apply to multiple data records. We consider it an AI-assisted UX paradigm because AI is used to partition (or batch) unlabeled data items into coherent groups<sup>1</sup> and user interface affordances allow users to take action on one or more items within the batch.

*3.2.1 Batch Labeling Interface.* Using a batching system, the overall labeling task proceeds by presenting a sequence of batches to the labeler, one batch at a time. The interface for labeling a batch is shown in Figure 2. Comparing it to the single-item interface in the system shown in Figure

<sup>1</sup>Informally, the contents of a group are coherent if all data items in the group are anticipated to have the same underlying label.

1, it includes mechanisms for selecting all items, toggling the selection of individual items in the batch on/off, and a count of the number of selected items. As in the single-item labeling process, the user can choose and confirm a label on the right. The "Select All" option is enabled by default, but the labeler can deselect items and apply a label to only those currently selected items. The batch stays on the screen, with labeled items becoming disabled, until all elements of the batch are labeled, at which point the system displays the next batch.

We focused on *fixed-size* batches for data labeling in order to limit the labelers' cognitive load, to create a more consistent labeling experience, and to remove the need for scrolling in the item area if there are more items in the batch than can fit on the screen at one time. For our purposes, we chose a batch size of 4 elements, balancing potential labeling efficiency gains against screen size constraints in the labeling interface and the cognitive load endured by the labeler.

**3.2.2 Nearest-Neighbor Batching.** In order to compute fixed-size coherent batches we use an unsupervised Nearest Neighbor (NN) inspired data partitioning algorithm. The algorithm produces batches based on computing an  $N \times N$  nearest neighbor matrix where  $N$  is the total number of data items. The  $j$ -th row in the matrix contains the  $N$  distance to all neighbors of data item  $j$  according to a similarity metric. Using the nearest neighbor matrix the algorithm iteratively constructs batches of size  $s$ . During each iteration a random un-batched item  $I$  is selected from the set of remaining un-batched data items. A batch is formed by including  $I$  and the  $(s - 1)$  closest un-batched neighbors. The algorithm continues until the set of un-batched items is exhausted. The NN batching algorithm relies on the *smoothness* assumption commonly applied in supervised and semi-supervised machine learning. This assumption states that points that are close to each other in the feature space are more likely to share a label.

Alternative data partitioning algorithms, such as K-Means [38] are not directly applicable to the data batching problem as they produce variable size partitions. The same limitation applies to general clustering algorithms, such as DBSCAN [16] which result in mixed size clusters and un-clustered outliers.

**3.2.3 Batch Quality Metrics.** To evaluate the quality of the batches we obtain with the NN algorithm, we consider the following set of metrics.

- *Homogeneity* [53] formalizes the notion of coherence and evaluates the quality of a dataset clustering relative to a known ground truth. A batching result satisfies homogeneity if all batches contain only data points which are members of a single class or label. Homogeneity values range between 0 and 1.
- *Mean Pair-Wise Cosine Similarity* within a batch provides a general measure of batch quality, with values ranging between 0 and 1. As per the *smoothness* assumption, batches with a higher degree of pairwise similarity, are more likely to share the same latent label.
- *Purity* [39] is a measure of the extent to which individual batches contain a single label, and is calculated in reference to a known ground truth. Purity is similar in nature to homogeneity, but applied on a per batch basis.

## 4 DATA AND BATCH CONSTRUCTION

For our research, we choose a short text classification problem as the underlying machine learning task for data labeling. Thus, the labeling task involves applying labels to a sequence of short text snippets (utterances). We ran three studies to investigate the relationship between labelers and the batching system. All experimental groups and details of the conditions involved in each study are listed in more detail in Table 1.



Group	Batching	Complexity	AI Quality	Feedback	Mitigation	Study
0	None	2 (Simple)	Baseline	None	None	1
1	Batching	2 (Simple)	Baseline	None	None	1
2	None	12 (Complex)	Baseline	None	None	1
3	Batching	12 (Complex)	Baseline	None	None	1, 2
4	Batching	12 (Complex)	Degradation	Present	None	3, 4
5	Batching	12 (Complex)	Baseline	Present	None	2, 3, 4
6	Batching	12 (Complex)	Perfect	Present	None	3
7	Batching	12 (Complex)	Baseline	Present	Rating	4
8	Batching	12 (Complex)	Baseline	Present	Machine-Teaching	4
9	Batching	12 (Complex)	Degradation	Present	Rating	4
10	Batching	12 (Complex)	Degradation	Present	Machine-Teaching	4

Table 1. Experimental manipulation for each group (15 participants per group, pre data-filtering). Study column indicates whether participants in the group were assigned to Study 1, Study 2, Study 3, and Study 4.

#### 4.1 Dataset / Data prep

The particular dataset chosen for these studies consists of short text utterances (9 words avg.), originally used to train an intent classification model for a customer service chatbot. To label the dataset, the labeler reads each utterance, and then selects a corresponding label (which is called an *intent* [70] in chatbot reinforcement learning paradigms). For example, the utterance “Can I request a replacement for my card online?” is correctly labeled with the intent “Replace Bank Card”.

The dataset consists of 240 utterances and twelve unique labels, each example having a single label, with a balanced distribution across labels in the set of utterances. The dataset was selected due to its general domain accessibility for non-specialists, and realistic use case. Table 2 shows the 12 original labels and their definitions, that belonged to two categories. For the purposes of our study we also derived a simpler categorical labeling of the data, where each example pertains to either “Banking” or “Telecommunication and Phone Services” (2 labels). For example, in the simple task (as pictured in Figure 1) participants can choose from “Banking”, whereas in the complex tasks “Banking” can be further subdivided into six categories: “Activate Bank Card”, “Cancel Bank Card”, “Fee Inquiry from Bank”, “Replace Bank Card”, “Report Missing Bank Card”, and “View Bank Routing Number”. The same applies to “Telecommunication and Phone Services.” We did this to keep the dataset constant across the various conditions.

Vocab. Size	Label	Definition
12	Activate Bank Card	Any request pertaining to the activation of a credit card or debit card
	Cancel Bank Card	Any request pertaining to the closing/canceling of a credit card or debit card
	Fee Inquiry from Bank	Any inquiry about a bank fee
	Replace Bank Card	Any request pertaining to replacing a credit card or debit card
	Report Missing Bank Card	Any request pertaining to the reporting of a missing/lost/stolen credit/debit card
	View Bank Routing Number	Any request to view routing number associated with a particular bank account/card
	Activate Phone/Device	Any request pertaining to the activation of a phone/sim or other telecommunications device
	Activate Roaming for Phone Device	Any request pertaining to the activation of roaming for a phone/device/sim
	Unlocking a Device/Phone	Any request pertaining to the unlocking for a phone/device/sim
	Changing/Keeping Phone Number	Any request pertaining to changing a current phone number or transferring a phone number between devices
	Return a Device/Phone	Any request pertaining to the returning of a phone/device/sim
	Swap a Device/Phone	Any request pertaining to the swapping/exchanging of a phone/device/sim
2	Banking	Any text that is pertaining to banking requests including credit card inquiries, banking fees, requests for data associated with banking
	Telecommunication and Phone Services	Any text that is related to requests around sim cards, phone services, roaming services and phone devices

Table 2. The labeling taxonomies and definitions used in the study.

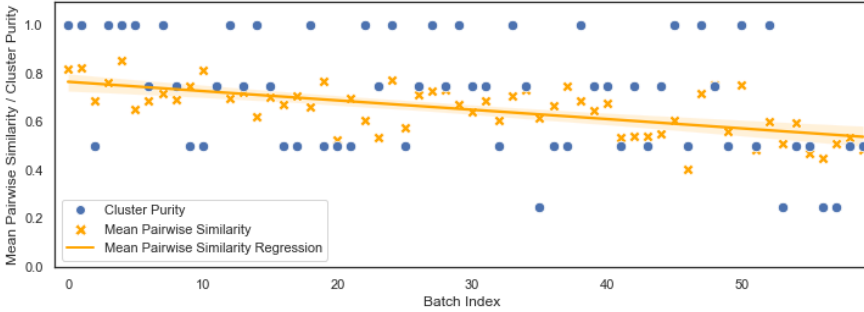


Fig. 3. A *batch chart* for Baseline AI batching. The X axis represents a sequence of batches, as presented to the labeler. Both purity and mean pairwise cosine similarity slightly degrade through the batches. For example, batch 30 has a purity of 1.0, indicating that all examples have the same label and the pairwise similarity of batch 30 is approximately 0.7.

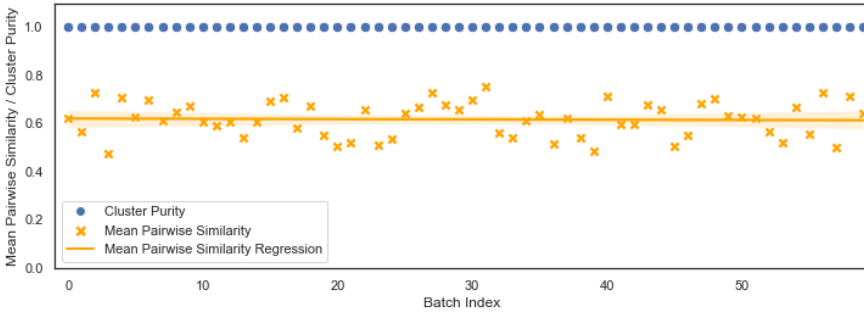


Fig. 4. Batch chart for *Perfect AI* batching applied to the study dataset. *Perfect AI* batching produces batches with a purity of 1.0. The fitted linear regression line shows that the mean pairwise cosine similarity is approximately constant across all batches.

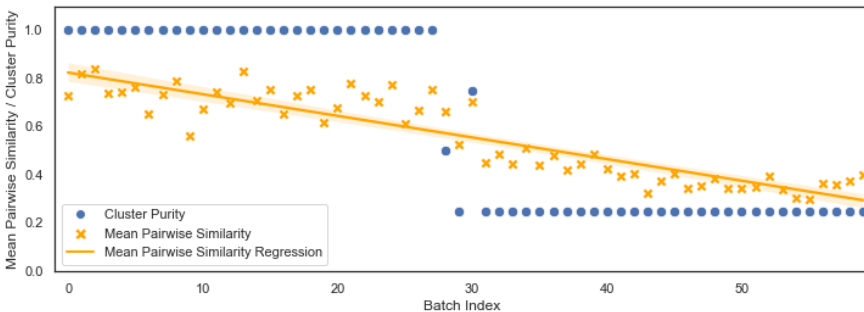


Fig. 5. Batch chart for Degradation AI batching applied to the study dataset.

## 4.2 Batch Construction

In order to apply the NN batch construction algorithm to our study data set we first encoded each text utterance in the dataset into a fixed size high-dimensional vector. To achieve this encoding we applied the Universal Sentence Encoder (USE) [10], a transfer-learning model capable of producing semantically rich vector representations from greater-than-word length text. USE produces a 512 component vector representation for each utterance. We applied the NN algorithm on the resulting vectors, using cosine similarity as the neighbor similarity metric.

In the remainder of the study we refer to the NN-based batch construction as *baseline AI batching*. To facilitate understanding and comparison among different batching algorithms we use *batch charts* as shown in Figure 3. Batch charts simultaneously plot the mean pairwise cosine similarity and purity within a batch and over a sequence of batches. The batch chart in Figure 3 provides a visual representation of Baseline AI batch construction, as applied to the study dataset. The figure illustrates how the mean pair-wise cosine similarity slowly degrades as a result of the greedy nature of the batch construction algorithm: during later batches, the closest, most similar neighbors may no longer be available for batch construction. Purity also slightly degrades throughout the batching. The average homogeneity score produced by Baseline AI batching on 3 different random orderings of the study dataset is 0.789. To study the effect of using AI-based batching for data labeling tasks, we are contrasting the baseline AI batching algorithm with two alternatives:

**4.2.1 Perfect AI Batching.** To simulate a *Perfect AI* batching algorithm, and to provide an upper limit to the quality of batches in such a system, we took advantage of the ground truth labels for our data set to construct batches that, by design, only include items with the same label. Both the average per batch purity and homogeneity of *Perfect AI* batching is 1.0. Figure 4 provides a visual representation of *Perfect AI* batching.

**4.2.2 Degradation AI Batching.** The converse to a *Perfect AI* batch is a worst-case batch where every item has a different label resulting in minimal purity and homogeneity. We constructed a degrading AI batching that starts out with *Perfect AI* batches and finishes with worst-case batches. We wanted to study an algorithm that might simulate some production scenarios, where the AI-Assistance algorithm starts strong but degrades over time (i.e. due to faulty online learning, etc). To do so, this "algorithm" initially replicates the batch purity/homogeneity of the *Perfect AI* - and then at a certain cut-off point will create increasingly more impure/nonhomogenous batches. To construct the Degrading AI batching we progressively grow the batching from both ends by taking turns between adding a new perfect batch in front and a new worst-case batch as we are growing from the end until all the data is contained in a batch. Because we were studying overreliance as a metric, we wanted to see whether users would accept the faulty recommendations given by an AI that started off perfect and degraded. Figure 5 displays the resulting batching for our study data set. The average homogeneity is 0.711.

## 5 METHODS

### 5.1 Participants

For Study 1, 2, 3, and 4 we recruited 165 total participants from Amazon Mechanical Turk [44]. Participating workers received a \$5.50 compensation based on an estimated work of 40 minutes for a projected US federal minimum wage (75%). The workers provided informed consent before completing the study. After the labeling task was complete, participants also answered demographic questions and questions about prior experience with Mechanical Turk as well as free-form questions explaining their impressions of the labeling system.

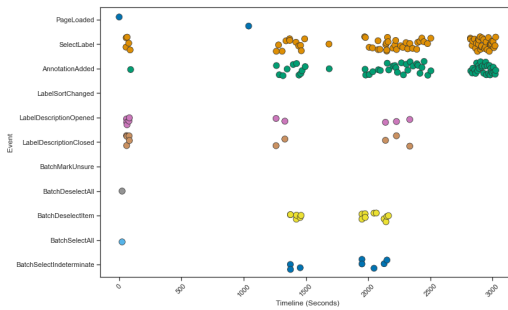


Fig. 6. Event timeline of a worker who restarted the task after a long pause, and then continued to take frequent breaks.

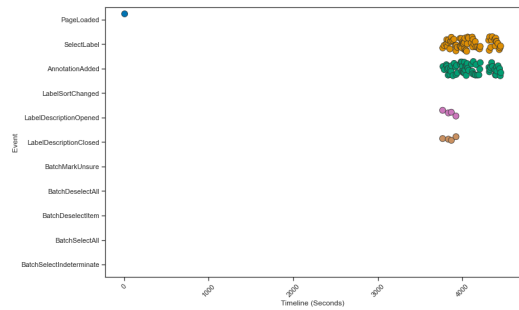


Fig. 7. Event timeline of a worker who started the task, then was inactive for over an hour, and finally completed without further breaks.

## 5.2 Data and Integrity Checks

**5.2.1 Integrity Checks.** We performed several integrity checks for our participants. Similar to prior studies deployed on Mechanical Turk [27], we excluded workers whose mean rating time was less than 5 seconds and did not deselect any item at any point during the labeling task. Deselecting is an indicator of paying attention to the task throughout and not merely clicking through every single batch, without reading the items. We also removed workers who had uniform responses ( $SD < 5$ ) in the survey responses. We also removed individuals from the study who fell outside of the  $mean \pm 2SD$  statistic for each of the dependent variables. This process left us with 156 participants. Examining the quality of the free-form responses at the end of the study showed that the manipulation checks were effective.

**5.2.2 Accounting for Breaks: Multi-tasking Users.** Prior work has shown that Mechanical Turkers multi-task, often taking breaks between tasks [33] or doing multiple HITs (Human Intelligent Tasks) simultaneously. While we offered a \$1.00 additional bonus to workers who completed the task efficiently, we observed that some workers took long breaks while completing tasks.

Anticipating the need for analyzing worker behavior, we instrumented the study interface to track pertinent events such as when a worker initially loaded into the labeling task, applied a label etc. Using this event stream we were able to construct a per second timeline for each worker as they performed the labeling task. Figures 6 and 7 show two examples of event timelines of workers who took long breaks during the labeling task.

To calculate actual labeling time, the period in which the worker was actively engaged, we removed individual trials that fell out of the  $mean \pm 2SD$  statistic for time spent on each individual trial.

## 5.3 Metrics Calculated

We measured the accuracy of labels (score out of 240 total labels), total time taken to complete labeling, user experience and trust towards the system. For each user, we collected the following dependent variables.

- **User Experience** We used the User Experience Questionnaire (Short Version) [34, 56], to measure perceived efficiency and enjoyment of the tool.
- **Duration** was calculated by measuring the time spent labeling in seconds on each labeling task. All user interactions were logged from the beginning of the labeling task to the end of the labeling used for Study 2, Study 3, and Study 4.

Demographic	
Mechanical Turk Experience	Less than 6 months (12.8 %), 6 months to a year (14%), 1-2 years (27%), 2-3 years (16.7%), More than 3 years (29.5%)
Language	English (82.7%), Tamil (10.9%), Portuguese (3.8%), Malayalam (0.06%), Hindi (1.3%), Chinese (0.06%)
Education	Middle School (5%), High School (22%), Bachelors or Higher (73%)
AI Experience	I closely follow AI-related news (23.1%), I have extensive experience in AI research and/or development (8.3%), I have heard about AI in the news, friends, or family (50%), I have never heard of AI (3.8%), I have some work experience and/or formal education related to AI (14.7%)

Table 3. Participant Demographics, N=156

- **Total Duration** was calculated by measuring the total time spent labeling in seconds. We used this metric in analysis for Study 1 since we were comparing 60 batching trials (4 items per trial; 240 total individual items) to 240 individual trials for non-batching conditions.
- **Accuracy** was measured as the percentage of individual items users labeled accurately.
- **Overreliance** When batching was involved across conditions (Study 2, 3, and 4), we calculate overreliance as the total of times a user followed the batched recommendations in assigning a batch the same label when the batch was incorrect.
- **Agreement** When batching was involved across conditions (Study 2, 3, and 4), we calculate agreement as the total of times a user followed the batched recommendations in assigning a batch the same label.

#### 5.4 Demographics and Prior Experience

We asked all participants involved in Study 1, Study 2, Study 3, and Study 4 about their demographics (primary language, education) and previous experience with Mechanical Turk HITs as well as with AI “What kind of exposure have you had to Artificial Intelligence (AI)” (1= I don’t know what machine learning is., 5= I have implemented a Machine Learning Algorithm). We present participant demographics in Table 3.

## 6 STUDY 1: BATCHING, TASK COMPLEXITY, ACCURACY, AND SPEED

In Study 1, we investigated the role of batching and task complexity and its impact on speed and accuracy of a labeling task. To evaluate the accuracy and speed in which users solve labeling tasks with AI-assisted batching we deployed the system on Mechanical Turk. The task begins with a consent form that describes procedures, risks, benefits, compensation, and participant rights. Participants are then randomly assigned to one of the following conditions (the experimental groups are further described in Table 1):

- Simple task with single-item (Group 0)
- Simple task with batch (multi-item) (Group 1)
- Complex task with single-item (Group 2)
- Complex task with batch (multi-item) (Group 3)

After completing the labeling task, users completed a post-survey that asks demographic questions, as well as open-ended questions about their experience with labeling. In Study 1, we are studying batching as a UX paradigm, keeping the batching algorithm constant across all conditions but manipulating the number of items recommended and number of labels in a task as a proxy for task complexity. This first study is a 2x2 between subjects study.

## 6.1 Experimental Manipulation

In Study 1, all participants interacted with the Baseline AI batching algorithm. The data and integrity checks described in Section 5.2 yielded a total of 36 participants. In Study 1, we addressed the following research questions:

**RQ1:** Does batching impact the outcome (accuracy score and time spent on task) of the labeling task?

**RQ2:** How does task complexity (vocabulary size) impact the outcome of data labeling?

## 6.2 Results

When presented with a batch labeling interface to help complete labeling tasks, do crowdworkers label more efficiently (accurately, faster)? We present our findings below.

**6.2.1 Labeling Performance: Accuracy.** We calculated a  $2 \times 2$  ANOVA to compare the main effects of batching (batching vs. none), task complexity (vocabulary size=2 vs. vocabulary size=12), and their interaction effect on accuracy. There were no significant differences for accuracy for batching or for vocabulary size. Neither was there any significant *Batch x Complexity* interaction.

**6.2.2 Labeling Performance: Duration.** We calculated a  $2 \times 2$  ANOVA to compare the main effects of batching (batching vs. none), task complexity (vocabulary size=2 vs. vocabulary size=12), and their interaction effect on duration. We found a significant effect of batching on duration ( $F_{1,32}=23.82$ ,  $p<0.001$ ). Tukey post hoc analysis showed that participants were significantly faster in Batching ( $628.14 \pm 215.35$ ) than in No-Batching conditions ( $1261.73 \pm 455.49$ ) ( $p < 0.05$ ). We also found a significant effect of task complexity on duration ( $F_{1,32} = 23.75$ ,  $p<0.001$ ). Tukey post-hoc analysis showed that participants in simple task conditions (vocabulary size=2) ( $738.70 \pm 317.17$ ) performed faster on tasks than participants in complex task conditions ( $1361.2 \pm 450.13$ ) ( $p < 0.05$ ).

In Study 1, we kept the algorithm used for batching constant across all conditions. The results from Study 1 demonstrate that batching has the potential to increase the speed of users in labeling tasks. We wanted to further explore the impact of quality of AI on batching performance and overreliance in Study 2. In Study 1, we also withheld information from participants about the batch recommendations being AI-driven and wanted to further investigate the role of this kind of feedback on performance, overreliance, and agreement with the batches.

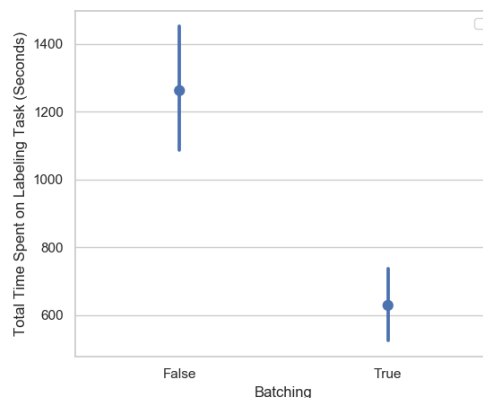
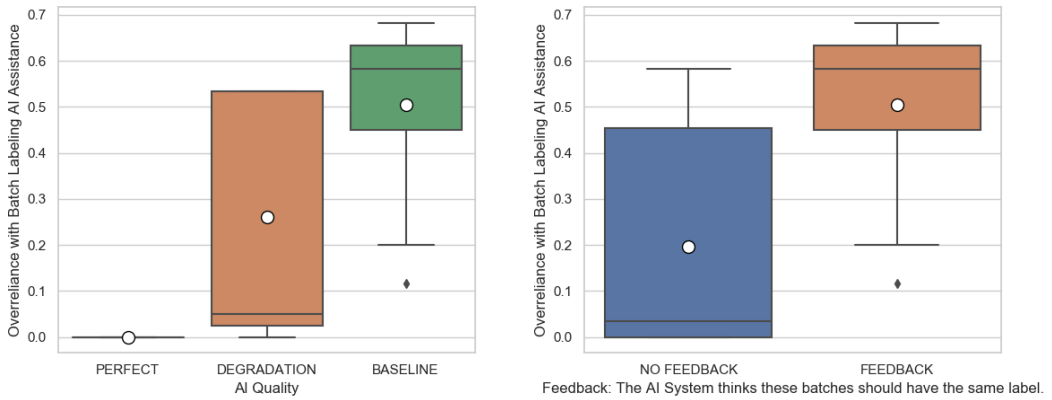


Fig. 8. Total time spent on task for conditions in which batching occurred and conditions in which batching did not occur in Study 1.



(a) Overreliance for AI quality from Study 3.

(b) Overreliance for Feedback from Study 2.

Fig. 9. Overreliance results from two Studies.

## 7 STUDY 2: INVESTIGATING FEEDBACK IN LABELING TASKS

The results of Study 1 show the benefits of a batch labeling interface (as compared to a single-labeling interface) for complex tasks in a crowdworker labeling task. For Study 2, we wanted to see if there would be differences if we told participants when they were interacting with an AI ("The AI System thinks these batches should have the same label") vs. when we did not give them feedback. We chose to study this in the context of the more complex task determined by Study 1 (12 labels, as opposed to 2), since the results of Study 1 showed the potential of batching for more complex tasks and prior research shows that people overrely or reach a point of complacency under more extreme workload conditions [32]. More specifically, in Study 2, we investigated the following research question: **RQ3**: Does telling users that the recommendations are from an AI impact the outcome (accuracy score, time spent on task, agreement with batches, and overreliance). Participants were randomly assigned to one of the following conditions (the experimental groups are further described in Table 1):

- Complex task, multi-item batches, Baseline AI, No Feedback (Group 3)
- Complex task, multi-item batches, Baseline AI, Feedback (Group 5)

**7.0.1 Accuracy, Duration, Agreement and Overreliance.** We ran a one way ANOVA for baseline participants who did not receive feedback and groups who did receive feedback (groups 3 and 5). We found no significant effect of feedback on accuracy ( $F_{1,21} = 0.72$  n.s.). We also ran a general linear mixed-effects model analysis of variance for duration for baseline participants who did not receive feedback and groups who did receive feedback (groups 3 and 5). We found no significant effect of feedback on duration. We also ran a one way ANOVA for baseline participants who did not receive feedback and groups who did receive feedback (groups 3 and 5) to investigate the impact of feedback on agreement with the AI and overreliance on the AI. We found no significant effect of feedback on agreement ( $F_{1,21} = 0.08$  n.s.). However, there was a significant effect of feedback on overreliance ( $F_{1,21} = 11.27$   $p < 0.01$ ). Tukey post-hoc analysis showed that users relied more on the AI when they were given feedback ( $0.51 \pm 0.19$ ) about the source of the batching ("The AI System thinks these batches should have the same label"), rather when they were not provided with any information ( $0.20 \pm 0.25$ )  $p < 0.01$ . Results can be seen in Figure 9b.

Given these findings, we moved forward with the investigation of **RQ4** in which we ran a between subjects study that investigated the impact of AI Quality (Baseline AI, Degradation AI, and Perfect AI) on speed, accuracy, agreement, and overreliance. All three conditions included the feedback ("The AI System thinks these batches should have the same label").

## 8 STUDY 3: INVESTIGATING QUALITY OF THE AI, AI AGREEMENT, AND OVERRELIANCE

We wanted to further investigate the impact the quality of the AI-driven batching can have on accuracy, agreement with batches, speed and overreliance. To compare our baseline AI batching to other AI conditions that might occur in production systems, we simulate two AI systems: *Perfect AI* and *Degradation AI*. As introduced in Section 4, *Perfect AI* batching produces perfect batch recommendations. In *Degradation AI* batching, the quality of the batch recommendations decreases over the course of batches. To address **RQ4**, we measure overreliance on the batching system as the percentage of incorrect recommendations by the batching system that a user assigns the same label.

### 8.1 Experimental Manipulation

In Study 3, we investigate the following question:

**RQ4:** How does the quality of the AI in batching impact the outcome (accuracy score, time spent on task, agreement with batches, and overreliance) of the batching task?

Participants were randomly assigned to one of the following conditions (the experimental groups are further described in Table 1):

- Complex task, multi-item batches, Degradation AI, Feedback (Group 4)
- Complex task, multi-item batches, Baseline AI, Feedback (Group 5)
- Complex task, multi-item batches, Perfect AI, Feedback (Group 6)

### 8.2 Results

**8.2.1 Labeling Performance: Accuracy and Duration.** When we consider user performance on labeling tasks, we measure two metrics: accuracy. We calculated a between-subjects ANOVA comparing the accuracy across the different batching conditions (*Perfect AI*, *Degradation AI*, *Baseline AI*). We found a significant effect of the type of AI used for accuracy ( $F_{2,40} = 3.45$ ,  $p < 0.05$ ). Tukey post-hoc analysis revealed significant differences between the *Baseline AI* ( $0.39 \pm 0.24$ ) and the *Perfect AI* ( $0.65 \pm 0.27$ ), with those in the *Perfect AI* condition labeling items more accurately than those in the baseline condition ( $p < 0.04$ ). For duration, a general linear mixed-effects model (a.k.a. mixed model) analysis of variance was used. AI quality was modeled as a fixed effect, while trials were nested within each subject and modeled as a random effect. We found no significant effect of the type of AI used on duration ( $F_{2,42.646} = 3.04$  n.s.).

**8.2.2 Agreement and Overreliance.** In Study 3, we were investigating how the quality of the AI impacts the degree to which users agree with the suggested items (giving all four items the same label) and overrely (giving all four items the same label when they should not have the same label). The ANOVA revealed significant differences for agreement for AI Quality ( $F_{2,40} = 6.90$ ,  $p < 0.01$ ). Tukey post-hoc analysis revealed significant differences for agreement between the *Degradation AI* ( $0.72 \pm 0.27$ ) and agreement for the *Perfect AI* ( $0.98 \pm 0.04$ )  $p < 0.01$ , showing that people agree significantly more with batches if the AI Quality is perfect.

Given our definition of "overreliance" in this context, it is impossible to overrely on batches recommended by the Perfect AI, since all *Perfect AI* batches are, by design, coherent (have the same label). We ran an ANOVA comparing overreliance across the three AI quality conditions ( $F_{2,40} =$



25.7,  $p < 0.001$ ). Tukey post-hoc analysis showed significant differences between the *Perfect AI* ( $0.0 \pm 0.0$ ) and Baseline AI ( $0.51 \pm 0.19$ )  $p < 0.001$ , significant differences between Baseline AI ( $0.51 \pm 0.19$ ) and the *Degradation AI* ( $0.26 \pm 0.26$ )  $p < 0.01$ , and significant differences between Degradation AI ( $0.26 \pm 0.26$ ) and the *Perfect AI* ( $0.0 \pm 0.0$ )  $p < 0.01$ . Results for differences in overreliance can be seen in Figure 9a. Study 3 demonstrated that when the AI is not perfect, as to be expected in a realistic scenario, labelers overrely (agree with the AI when they should not). We also find that users overrelied more on the *Baseline AI* than the *Degradation AI*. These results led us to investigate mitigation of overreliance in our fourth study. Given that people overrely when presented with *Baseline AI* or *Degradation AI*, what can the kind of information can we provide in the interface to mitigate their overreliance?

## 9 STUDY 4: INVESTIGATING OVERRELIANCE MITIGATION

In Study 4, we investigate the mitigation of overreliance by addressing the following research questions. We evaluated the effectiveness of ameliorating overreliance by implementing two mitigations in Study 4 as seen in Figures 10 and 11.

**RQ5:** Can we mitigate overreliance on AI-assisted batching by asking users to rate the batches suggested by the AI system?

**RQ6:** Can we mitigate overreliance on AI-assisted batching by signaling to users that their responses will help improve the AI's batch recommendations in the future.

Participants were randomly assigned to one of the following conditions (the experimental groups are further described in Table 1):

- Complex task, multi-item batch, Baseline AI, Feedback, No Mitigation (Group 5)
- Complex task, multi-item batch, Baseline AI, Feedback, Rating (Group 7)
- Complex task, multi-item batch, Baseline AI, Feedback, Machine-Teaching (Group 8)
- Complex task, multi-item batch, Degradation AI, Feedback, Rating (Group 9)
- Complex task, multi-item batch, Degradation AI, Feedback, Machine Teaching (Group 10)
- Complex task, multi-item batch, Degradation AI, Feedback, No Mitigation (Group 4)

### 9.1 Mitigation 1: Machine Teaching

In this mitigation, we informed users that they were collaborating with the AI system to recommend better batches. Users were told, "Your labels help train the AI system to recommend better batches." This mitigation was meant to signal that all the batches were not necessarily perfect and that users were working with the AI to ultimately help improve the batches for future uses of the system. Consistent with prior work [50], this mitigation meant to frame the relationship between the AI-driven batching system and the user as a collaborative one, in which the labeler works with the recommended batches to improve them for future users.


### 9.2 Mitigation 2: Rating

In the Rating Mitigation, users were informed that the quality of batches would be rated at the completion of the study, signaling that the AI is not perfect. Users were notified that "You will be asked to rate the AI on the quality of these batches in a survey".

### 9.3 Experimental manipulation

For Study 4, the attention check described in Section 5.2 yielded 77 participants. Participants were assigned to either the Baseline AI condition, the Degradation AI condition, or the Mitigation condition (None, Rating, Machine-Teaching).

0/60 labeled


 **The AI system thinks these items should have the same label**  
Your labels help train the AI system to recommend better batches.

Current Items  Select All Items 4/4

	<input checked="" type="checkbox"/> Item 0	:	<input checked="" type="checkbox"/> Item 1	:	<input checked="" type="checkbox"/> Item 2	:	<input checked="" type="checkbox"/> Item 3	:
example	I need assistance as I've lost my credit card		What steps can I take when I lose my credit card?		I have lost my credit card and I want to deactivate it		I would like to know what to do if I lose my credit card	

Fig. 10. Machine Teaching Mitigation pictured above in which users are notified that their responses help train the AI to give better batches.

0/60 labeled

 **The AI system thinks these items should have the same label**  
You will be asked to rate the AI on the quality of these batches in a survey.

Current Items  Select All Items 4/4

	<input checked="" type="checkbox"/> Item 0	:	<input checked="" type="checkbox"/> Item 1	:	<input checked="" type="checkbox"/> Item 2	:	<input checked="" type="checkbox"/> Item 3	:
example	Where can I find my bank routing numbers for money transfer?		How to search bank routing numbers?		How do I know the routing number?		How to locate bank routing number?	

Fig. 11. Rating Mitigation pictured above in which users are notified that they will be rating the quality of the batches in a survey at the end of the study.

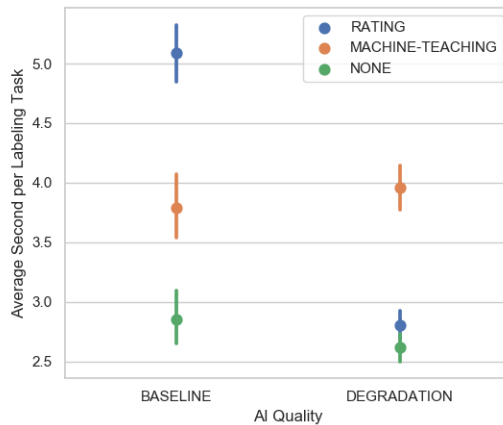


Fig. 12. Duration for AI Quality and Mitigation measures in Study 4.

### 9.4 Results

9.4.1 Accuracy and Duration. We first ran a 2x3 ANOVA to compare the main effects of AI Quality (Baseline AI and Degradation AI) and Mitigation type (Rating vs. Machine-Teaching vs. No Mitigation), and their interaction effect on accuracy and found no statistically significant effect of

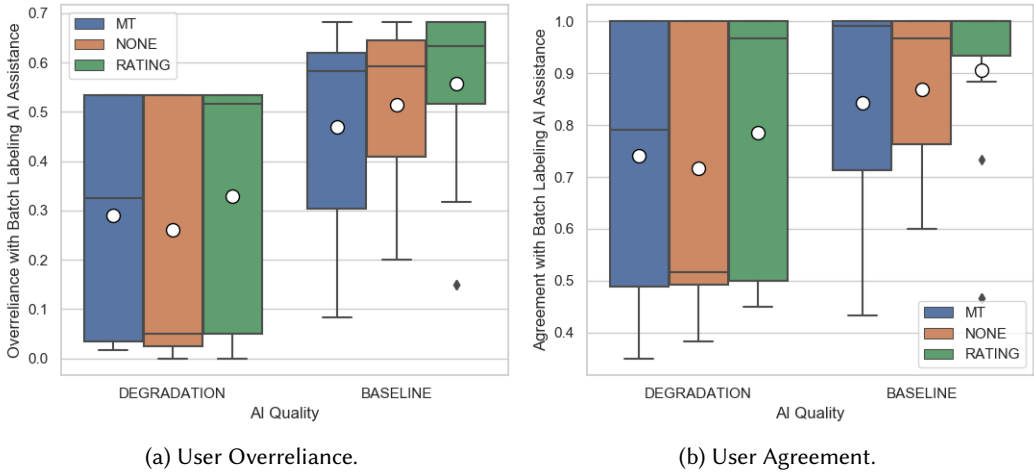


Fig. 13. Overreliance and Agreement Results from Study 4, in which we investigated the effect of the three mitigations (Machine-Teaching, Rating, None) across different AI qualities (Degradation, Baseline) on user agreement and overreliance.

mitigation type on accuracy ( $F_{2,71} = 0.23$  n.s.), no significant effect of AI quality on accuracy ( $F_{1,71} = 0.50$  n.s.) and no significant interaction effect ( $F_{2,71} = 1.124$  n.s.).

For duration, a general linear mixed-effects model (a.k.a. mixed model) analysis of variance was used. The mitigation condition (Rating, Machine-teaching, No Mitigation) and the AI type (baseline, degradation) were modeled as fixed effects, while trials were nested within each subject and modeled as a random effect. We found a significant interaction effect between AI quality and mitigation type ( $F_{2,74.382} = 3.68$   $p < 0.05$ ), and no significant main effect for mitigation type ( $F_{2,74.382} = 2.00$  n.s.) or AI quality ( $F_{2,74.374} = 0.45$  n.s.). Post-hoc tukey analysis revealed significant differences ( $p = 0.045$ ) between Rating Mitigation  $\times$  Baseline AI ( $5.09 \pm 8.19$ ) and No Mitigation  $\times$  Degradation AI ( $2.61 \pm 4.95$ ) and marginally statistically significant differences ( $p = 0.072$ ) between Rating Mitigation  $\times$  Baseline AI ( $5.09 \pm 8.19$ ) and Rating Mitigation  $\times$  Degradation AI ( $2.80 \pm 4.77$ ). Figure 12 shows results of the experiment. The significant time difference between the Rating Mitigation and other groups without a decrease in accuracy suggests that users are thinking more about their decision making when they are presented with the Rating Mitigation.

**9.4.2 Agreement and Overreliance.** We ran a two-way ANOVA comparing overreliance across the two AI quality conditions (baseline, degradation) and three mitigations (Machine-Teaching, Rating, None). We found no statistical differences for mitigation type ( $F_{2,71} = 0.622$  n.s.) or interaction effects ( $F_{2,71} = 0.151$ , n.s.). We did find a significant effect for AI Quality ( $F_{1,71} = 17.24$ ,  $p < 0.001$ ), supporting our findings from Study 3 that AI Quality has an impact on overreliance. Tukey post-hoc analysis showed significant differences between the Degradation AI ( $0.29 \pm 0.25$ ) and Baseline AI ( $0.52 \pm 0.20$ )  $p < 0.001$ . We also ran an ANOVA to investigate user agreement with the batching system. We found there was also a statistically significant main effect of AI quality  $F_{1,71} = 5.078$   $p < 0.001$  on agreement. There was no statistically significant main effect mitigation on agreement ( $F_{2,71} = 0.48$  n.s.) and no interaction effect ( $F_{2,71} = 0.06$  n.s.). These result were surprising since our mitigations were designed to decrease overreliance and we saw no effect on overreliance.

## 10 USER EXPERIENCE

For each of the studies, we ran a survey to investigate the effect of our conditions on user experience. For each study we ran an ANOVA to understand the effects of the various conditions on the user experience. In Study 1, we investigated the impact of batching and task complexity on user experience. In Study 2, we investigate the effect of feedback on user experience. In Study 3, we investigated the effect of AI Quality on user experience. In Study 4, we investigated the effect of AI Quality and mitigation types on user experience. These user experience measures were an average of eight questions asked in the post survey on a 1 to 7 likert scale about the user's experience on the platforms [34, 56]. We did not find any statistically significant effects with respect to user experience. However, we followed up with open-ended questions that give us insight into user reactions to the labeling assistance tool. We describe these findings in the next section.

## 11 OPEN-ENDED QUESTIONS

At the end of the labeling task, we asked each participant five different open-ended questions, "Please describe why you think the recommendations in the labeling tool were helpful/not helpful", "How would you describe your experience with the labeling tool?", "What did you like best about your experience with the labeling tool?", "What did you like least about your experience with the labeling tool", and "What would you change about the labeling tool?" Two of the co-authors looked at responses for each group (referenced in Table 1) and extracted themes. We list those results below.

**11.0.1 Users Noticed the Quality of the AI.** Across the groups, users commented on the quality of the AI. Particularly for participants who were given the *Degradation AI* (Group 4, 9, and 10 in Table 1), users noticed the quality of the batches decreasing after the first 30 recommendations.

*The first 30 sets were grouped well but the last 30 were almost all completely unrelated* (Participant 73, Degradation AI, Group 4).

*I think they were helpful till about half way through but the messages got vague or used uncommon language. I think the AI was good at certain words but it needs to learn a bigger vocab* (Participant 67, Degradation AI, Group 4)

Users also commented on the quality of AI for the Perfect conditions. Particularly for the Perfect AI, users applauded the system.

*The work was already pretty much done for me* (Participant 56, Perfect AI, Group 6).

*The system was more advanced than I thought.* (Participant 61, Perfect AI, Group 6)

*The recommendations were mostly related with the label so it was very helpful. Only in few cases I had to change it* (Participant 123, Baseline AI, Group 7)

**11.0.2 Complex Tasks Confuse Users.** As Study 1 demonstrated, users take longer on complex tasks (vocabulary=12). The feedback from our open ended responses gave further evidence that labels confuse individuals.

*"I think there were too many categories to choose from and that just leads to confusion"* (Participant 87, Baseline AI, Group 5).

**11.0.3 Batching improves User Experience Except When it Does Not.** While we did not see any statistical differences in the user experience metrics we collected. Some comments across the different groups that employed batching showed that users felt that batching improved their overall experience.

*At first the idea of doing 60 groups in a row seemed like a lot but it ended up being a lot easier and more efficient than I expected (Participant 50, Baseline AI, Group 3).*

However, users commented on the frustrating user experience of un-checking and rechecking items when presented with bad batches in the Degradation AI. Other users complained about the large number of categories.

## 12 DISCUSSION

Our results demonstrate that batching can improve time and accuracy. This outcome relies partly on the complexity of the task and the quality of the AI, since our results show that users have to do more work when the batches recommended for one label are not coherent. We present design recommendations for labeler-AI-batching system interaction emerging from our results and experience in building and studying the system.

### 12.1 Design Implications

In the real world, the people who create tasks for labelers do not know the ground truth of the labels. The purpose of the labeling task is to collect the ground truth. We keep this in mind as we make recommendations for our dataset. Certain metrics like mean pair-wise similarity can be calculated in the absence of ground truth knowledge, while metrics like homogeneity and purity cannot. In the real world, individuals assigning these labeling tasks would not have access to metrics like homogeneity and purity, i.e. further challenges arise for them in understanding how well a batching algorithm works and, correspondingly, what the impact will be on the performance of the labelers. In some citizen science projects, the administrators (professional scientists) sometimes check the quality of their volunteer labelers by inserting test items *which have known ground-truth answers* into the labelers' workflow [69]. If needed, it may be possible use a similar approach to obtain estimates of homogeneity and purity in real-world batching experiences.

As we discussed earlier, labeling usually occurs as part of the team's work in data science [35, 43, 48]. Teams frequently engage in a division-of-labor according to specialized roles and skills [19, 30, 40, 68, 69]. This paper can help with that division-of-labor, especially in view of increasing AI capabilities that can shift the balance of initiative between human and AI [14, 26, 58, 63]. Our results can further advise administrators and data scientists as they set up the tasks of labelers [76]. When is batching helpful? When is a simpler user interface more beneficial? We will provide summary answers to those questions in this section.

**12.1.1 Overreliance.** This study design allowed us to operationalize overreliance [21, 23, 47, 52] by defining it as the degree to which users agree with a batch recommendation from the AI but are incorrect (e.g., similar to [58]). We explored different AI qualities in Study 3 (degradation, baseline, perfect) and found significant differences in quality not just when the AI is perfect, but also between the degradation AI and the baseline AI. In fact, users overrelied on the baseline AI almost twice as much as they did for the degradation AI. Our results show that the quality of the AI impacts the degree to which users agree with batch agreements, showing that users notice the quality of the AI when making decisions. Our open-ended responses showed that users noticed the quality of the AI began to decrease.

This might explain why users did not overrely on the Degradation AI as much as they did on the Baseline in which the quality of batches remained constant throughout the labeling task. We found that users who received the feedback, *"The AI System thinks these batches should have the same label"* were more likely to overrely (agree with the AI's recommendations) and be incorrect twice as much as when they did not see this feedback from the system. This distinction - explicitly telling users that the recommendations were AI generated - resulted in this overreliance (see [45] for

similar evidence of human overreliance on AI). Prior work on automation bias in clinical decision support systems found several mitigators of overreliance, including position of advice on the screen, confidence levels attached to the automated recommendations and providing information rather than recommending [21]. The position and wording of our feedback (authoritativeness of the statement) may have impacted the degree to which users overrelied on the system.

*12.1.2 Select-All functionality should only be used if there is certainty around the similarity of batches.* In all of our conditions, all items in batches recommended to individuals were selected by default. When items in the batch were not coherent (did not all have the same label), users expressed frustration over having to “uncheck” items and do more work, particularly in the Degradation AI condition.

*After the first 30 sets the inquires were all unrelated and it was tedious to uncheck then check the relevant inquiries and assign labels (Participant 73, Degradation AI, Group 4).*

If metrics like pairwise-similarity and other features (n-grams) used in the batching algorithm indicate high levels of similarity, Select-All can be used. However, if items are not similar, having all items pre-selected by default leads to less optimal user experience for the user. Alternatively, if the uncertainty about items is known upfront, uncertain items can be displayed unselected per default (e.g. in our NN algorithm, a similarity threshold could be used).

*12.1.3 For batching systems with varying batch sizes, the number of items recommended in a batch should depend on the similarity of items in the batch.* In our labeling tool, the number of items displayed on a page was fixed to four per batch. In conditions where the AI was perfect, all four items were to be assigned the same label. Users in this condition commented on how helpful they believed the system to be: *“It removed whatever doubt I might have had about labeling categories”* (Participant 56, Perfect AI, Group 6). However, in conditions in which the AI was not perfect, four items seemed like too many items to users. *“I would give some kind of guideline for times when the results were mixed. Perhaps reduce the number of results at times too. Four seems like a lot”* (Participant 88, Baseline AI, Group 5). In future designs, the number of items per batch could rely on metrics of similarity that can be extracted from the data. For example, if two items (Item 1 and Item 2) in a dataset are more similar than when a third item is added (Item 3), only Item 1 and Item 2 should be presented to the user as a batch. It may also be useful to allow labelers to adjust the size of their batches as needed, since cognitive load might vary from one labeler to another. Some participants complained that four is too much, while others did not mind. Giving users this sense of agency can potentially improve the user experience for the labeler. Prior work has shown the benefits of adaptable interfaces in mitigation overreliance [41]. In an alternate user experience design, we can also envision the user loading more items dynamically (e.g. with a “More Like This” button) and, based on the users ongoing learning of the data characteristics, the UI will display more items in decreasing similarity. Eventually, the items might reach a cut-off and require a different label at which point the user could switch over to a “different label” batch loading.

*12.1.4 Feedback Impacts Performance.* The results from Study 2 showed that participants who were presented with the following feedback (*“The AI System thinks these batches should have the same label”*) overrelied more than those who did not see the feedback, suggesting that when participants think the batches are AI-recommended, they rely on the batches more than if users were told they were not recommended by an AI. We suggest that designers of labeler-AI-batching system interaction should consider the balance between avoiding user overreliance on the system and user time and accuracy by examining how design choices can affect how users perceive how the batches are made. For example, do labelers perceive the batches to be AI-driven? Do labelers perceive their

selection is helping to train a less mature AI? This consideration can assist in designing systems that discourage overreliance so that labelers can exert critical judgement over the batch suggestions and create higher quality labeled data. It is also possible that the particular Machine Teaching mitigations we employed were less effective than other kinds of mitigations that cause a labeler to reflect mindfully on their decisions, which may be useful in reducing overreliance. For example, feedback could be given to show how often the labeler seemed to be relying on the batch or how quickly they were labeling. This is an open problem for future study.

In Study 4, we saw that users took a much longer time on average on a labeling task in the Rating Mitigation  $\times$  Baseline AI condition. One potential explanation for this is that users felt the need to concentrate more on the batches since they would be asked to rate the system upon completion of the labeling task. We did not observe this same effect for users in the Rating Mitigation  $\times$  Degradation AI condition. One potential reason for this is that the *Degradation AI* starts with perfect batches and degrades over time, potentially decreasing the amount of time, while the *Baseline AI* has the same quality recommendations throughout, sometimes imperfect. For example, a participant in the Baseline condition said, “*I would give some kind of guideline for times when the results were mixed. Perhaps reduce the number of results at times too. Four seems like a lot*” (Participant 88, Baseline AI, Group 5), suggesting that users notice the quality of the suggestions in the *Baseline AI*. If they begin to doubt the quality early on, they may on average spend longer time on each labeling task, especially since they are required to rate the AI at the end of the study. While we saw increased duration for one of our mitigation conditions (Rating Mitigation) and no changes in accuracy, future work can further investigate other mitigation strategies to achieve improvement in both accuracy and duration. The increased duration in the Rating Mitigation condition shows that users can spend a longer time on a task without necessarily being more accurate. Future work can attempt to communicate the collaborative nature of the task (between the human and the AI-driven assisted labeling tooling) through alternative ways to increase accuracy, decrease duration, and decrease overreliance.

*12.1.5 Label Suggestions May Help: Perfect Batches don't Yield Perfect Scores.* In Study 3, we found that there was an 98% agreement rate by labelers with the batches for the *Perfect AI* conditions. However, the accuracy score was only 65%. This means that in a scenario in which all the items in a batch had one label, individuals still mislabeled the content, assigning the same labels to the batch but the incorrect label. As we reported in Section 11.0.1, one participant even commented, “The work was already pretty much done for me.” (Participant 60, Perfect AI, Group 6). This kind of attitude points to a different type of overreliance compared to the one we measured. A way to address this kind of interaction is to not only suggest items that are likely to have the same label, but suggest the labels themselves for items - but only in cases in which the system is sufficiently confident about those label suggestions.

## 12.2 Ethical Implications for Labeling

We must acknowledge that many crowdworkers on these platforms are not one-time users, but also individuals who earn their income through these platforms [71]. In the age of crowdworking platforms, scholars have categorized two classes of Mechanical Turk Stakeholders: one that outsources small repetitive tasks and one that is on the receiving end doing these mundane tasks [55]. This taxonomy results in a dystopian situation where people do not know if a machine or a human is handling the mundane task at hand. With any platform involving crowdworkers, such dynamics must be considered. The goal of the labeling system is not to only benefit the individuals creating tasks, but also to improve the user experience of labeling. By speeding up the process through batching, all of the individuals in the crowdworking pipeline (both those assigning tasks

as well those doing them) benefit. If designed correctly, crowdworkers can label more efficiently. In medical domains and others that require SME input for labeling, SME's time is valuable, limited, and expensive. In scenarios in which labeling is required from subject matter experts, our results show that batch-labeling can speed up the number of items individuals are able to label correctly in a set amount of time, leading to more robust, accurate AI systems as a result.

### 13 LIMITATIONS AND FUTURE WORK

The batch labeling system we introduce is meant to help labelers to label text-based unstructured data in different contexts by framing the labelling task as a labeler-machine collaborative effort. Using this framing, the resulting design implications may apply to cross-media labelling situations, where a labeler can also work together with an AI-labeling assistive system. Previous works highlight that repurposing labeling tools designed for one medium to another is viable [9, 18, 73] and to fortify the applicability of our findings in the design of cross-media labelling tasks we recommend the following future work.

First our study was conducted on one dataset that consisted of short structured data. Future work should explore how the experimental conditions in this study impact different data sets that are structured differently than the dataset we explored. Secondly, we acknowledge the limitations for recruiting participants from Mechanical Turk. Crowdworkers recruited from Mechanical Turk are not subject matter experts. Results and respective design implications may differ for similar studies for subject matter experts across various domains. Future work should investigate batching accuracy, speed, agreement and overreliance for Subject Matter Experts.

Working with crowdworkers has limitations in terms of the quality of the data collected [42]. While the accuracy in our studies is low, our goal is not to train a model. Rather, we are trying to understand factors that could influence the labeling practices that are part of training a model. In principle, we wanted to choose a task that is difficult (for this particular set of labelers), exactly because that level of difficulty may expose factors that would be less apparent in an easier task. Lastly, while we did not find statistically significant results for our mitigations, we found that feedback has the potential to impact outcome of the labeling task (i.e. duration in Study 3). Future research should explore other mitigations for overreliance and their impact on user performance.

### 14 CONCLUSION

In this paper, we conducted four studies to investigate the role of batching, task complexity, and AI Quality on labeling performance. Our studies demonstrate that "batch-labeling", an AI-assisted UX paradigm leads to labelers completing tasks more accurately (given the quality of the AI) and faster. By manipulating the AI Quality (Degradation AI, Perfect AI, Baseline AI) across conditions we also find that the AI Quality impacts the degree to which users agree with batching recommendations and overreliance on batching recommendations. More research is needed to design and investigate overreliance mitigations in a labeler-AI-batching system interaction context. Our findings offer implications for the design of batching labeling systems and for work practices focusing on labeler-AI-batching system interaction.

### REFERENCES

- [1] Mahmoud J Abu Ghali, Abdullah Abu Ayyad, Samy S Abu-Naser, and Mousa Abu Laban. 2018. An intelligent tutoring system for teaching english grammar. (2018).
- [2] Janna Anderson, Lee Rainie, and Alex Luchsinger. 2018. Artificial intelligence and the future of humans. *Pew Research Center* (2018).
- [3] Josh Andres, Christine T Wolf, Sergio Cabrero Barros, Erick Oduor, Rahul Nair, Alexander Kjærsum, Anders Bech Tharsgaard, and Bo Schwartz Madsen. 2020. Scenario-based XAI for Humanitarian Aid Forecasting. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.



- [4] Yukino Baba and Hisashi Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 554–562.
- [5] Yulia Badryzlova, Natalia Shekhtman, Yekaterina Isaeva, and Ruslan Kerimov. 2013. Annotating a Russian corpus of conceptual metaphor: A bottom-up approach. In *Proceedings of the First Workshop on Metaphor in NLP*. 77–86.
- [6] Victoria A Banks, Katherine L Plant, and Neville A Stanton. 2019. Driving aviation forward; contrasting driving automation and aviation automation. *Theoretical issues in ergonomics science* 20, 3 (2019), 250–264.
- [7] Anthony Brew, Derek Greene, and Pádraig Cunningham. 2010. The interaction between supervised learning and crowdsourcing. In *NIPS workshop on computational social science and the wisdom of crowds*.
- [8] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [9] Andrew Caines, Christian Bentz, Calbert Graham, Tim Polzehl, and Paula Buttery. 2016. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the crowdis corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2145–2152.
- [10] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Lintiac, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv preprint abs/1803.11175* (2018). arXiv:1803.11175 <http://arxiv.org/abs/1803.11175>
- [11] Alan Chamberlain, Alessio Malizia, and David De Roure. 2017. An agent on my shoulder: AI, privacy and the application of human-like computing technologies to music creation. (2017).
- [12] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.
- [13] Meltem Demirkus, James J Clark, and Tal Arbel. 2014. Robust semi-automatic head pose labeling for real-world face video sequences. *Multimedia Tools and Applications* 70, 1 (2014), 495–523.
- [14] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 628–635.
- [15] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
- [16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.
- [17] Julien Fauqueur, Ashok Thillaisundara, and Theodosia Togia. 2019. Constructing large scale biomedical knowledge bases from scratch with rapid annotation of interpretable patterns. *arXiv preprint arXiv:1907.01417* (2019).
- [18] Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 80–88.
- [19] Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.
- [20] J Glanz, M Suhartono, and H Beech. 2018. In Indonesia Lion Air crash, black box data reveal pilots' struggle to regain control. *The New York Times* (2018).
- [21] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.
- [22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [23] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
- [24] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [25] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [26] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.
- [27] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 239.
- [28] Kaggle. 2017. 2017 Kaggle ML & DS Survey. <https://www.kaggle.com/kaggle/kaggle-survey-2017>

- [29] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 1941–1944.
- [30] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 96–107.
- [31] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 487.
- [32] Ellen J Langer. 1989. Minding matters: The consequences of mindlessness–mindfulness. In *Advances in experimental social psychology*. Vol. 22. Elsevier, 137–173.
- [33] Laura Lascau, Sandy JJ Gould, Anna L Cox, Elizaveta Karmannaya, and Duncan P Brumby. 2019. Monotasking or Multitasking: Designing for Crowdworkers’ Preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [34] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.
- [35] George Lawton. 2018. The nine roles you need on your data science research team. TechTarget. <https://searchcio.techtarget.com/news/252445605/The-nine-roles-you-need-on-your-data-science-research-team>.
- [36] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
- [37] Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J Quinn. 2016. Crowdsourcing high quality labels with a tight budget. In *Proceedings of the ninth acm international conference on web search and data mining*. 237–246.
- [38] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.
- [39] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.
- [40] Kate Matsudaira. 2015. The science of managing data science. *Queue* 13, 4 (2015), 30.
- [41] Christopher A Miller, Harry Funk, Robert Goldman, John Meisner, and Peggy Wu. 2005. Implications of adaptive vs. adaptable UIs on decision making: Why “automated adaptiveness” is not always the right answer. In *Proceedings of the 1st international conference on augmented cognition*. 22–27.
- [42] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.
- [43] Michael Muller, Melanie Feinberg, Timothy George, Steven J Jackson, Bonnie E John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, W15.
- [44] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.
- [45] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.
- [46] Raja Parasuraman, Robert Molloy, and Indramani L Singh. 1993. Performance consequences of automation-induced ‘complacency’. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23.
- [47] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
- [48] DJ Patil. 2011. *Building data science teams*. " O’Reilly Media, Inc."
- [49] Claudio Pinhanez. 2019. Machine Teaching by Domain Experts: Towards More Humane, Inclusive, and Intelligent Machine Learning Systems. *arXiv preprint arXiv:1908.08931* (2019).
- [50] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* (2020), 1–39.
- [51] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal* 29, 2 (2020), 709–730.
- [52] Victor Andrew Riley. 1994. *Human use of automation*. Ph.D. Dissertation. ProQuest Information & Learning.
- [53] Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 410–420.
- [54] Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. 2020. Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

*Systems*. 1–13.

- [55] Florian Alexander Schmidt. 2017. *Crowd Design: From Tools for Empowerment to Platform Capitalism*. Birkhäuser.
- [56] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *IJLMAI* 4, 6 (2017), 103–108.
- [57] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
- [58] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (2020), 495–504.
- [59] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 102–107.
- [60] Jack Stewart. 2017. Don't freak over Boeing's self-flying plane—robots already run the skies'. *Wired*, (online, 6 September 2017) (2017).
- [61] Catherine Stinson. 2018. *Healthy Data: Policy solutions for big data and AI innovation in health*. Mowat Centre for Policy Innovation.
- [62] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- [63] Lucy A Suchman. 1990. What is human-machine interaction. *Cognition, computing, and cooperation* (1990), 25–55.
- [64] S Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 538.
- [65] Szilárd Vajda, Yves Rangoni, and Hubert Cecotti. 2015. Semi-automatic ground truth generation using unsupervised clustering and limited manual labeling: Application to handwritten character recognition. *Pattern recognition letters* 58 (2015), 23–28.
- [66] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [67] Daniel Karl I Weidele, Justin D Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. 2020. AutoAIViz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 308–312.
- [68] Andrea Wiggins and Kevin Crowston. 2010. Developing a conceptual model of virtual organisations for citizen science. *International Journal of Organisational Design and Engineering* 1, 1-2 (2010), 148–162.
- [69] Andrea Wiggins and Kevin Crowston. 2011. From conservation to crowdsourcing: A typology of citizen science. In *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.
- [70] Jason D Williams, Nopal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. In *Natural Language Dialog Systems and Intelligent Assistants*. Springer, 1–13.
- [71] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81.
- [72] Christine Wolf and Jeanette Blomberg. 2019. Evaluating the Promise of Human-Algorithm Collaborations in Everyday Work Practices. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 143 (Nov. 2019), 23 pages. <https://doi.org/10.1145/3359245>
- [73] Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2013. Brat2BioC: conversion tool between brat and BioC. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, Vol. 1. 46–53.
- [74] Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 1–6.
- [75] Ming Yin, Jennifer Wortman Vaughan, and Hanna 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, 279.
- [76] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

Received June 2020; revised October 2020; accepted December 2020