# A Human-Centered View of Continual Learning: Understanding Interactions, Teaching Patterns, and Perceptions of Human Users Toward a Continual Learning Robot in Repeated Interactions

ALI AYUB, ZACHARY DE FRANCESCO, JAINISH MEHTA, and
KHALED YAAKOUB AGHA, University of Waterloo, Waterloo, ON, Canada
PATRICK HOLTHAUS, University of Hertfordshire, Hertfordshire, UK
CHRYSTOPHER L. NEHANIV and KERSTIN DAUTENHAHN, University of Waterloo, Waterloo, ON, Canada

Continual learning (CL) has emerged as an important avenue of research in recent years, at the intersection of Machine Learning (ML) and Human–Robot Interaction (HRI), to allow robots to continually learn in their environments over long-term interactions with humans. Most research in CL, however, has been *robot-centered* to develop CL algorithms that can quickly learn new information on systematically collected static datasets. In this article, we take a *human-centered* approach to CL, to understand how humans interact with, teach, and perceive CL robots over the long term, and if there are variations in their teaching styles. We developed a socially guided CL system that integrates CL models for object recognition with a mobile manipulator robot and allows humans to directly teach and test the robot in real time over multiple sessions. We conducted an in-person study with 60 participants who interacted with the CL robot in 300 sessions with 5 sessions per participant. In this between-participant study, we used three different CL models deployed on a mobile manipulator robot. An extensive qualitative and quantitative analysis of the data collected in the study shows that there is significant variation among the teaching styles of individual users indicating the need for personalized adaptation to their distinct teaching styles. Our analysis shows that the constrained experimental setups that have been widely used to test most CL models are not adequate, as real users interact with and teach CL robots in a variety of ways. Finally, our analysis shows that although users have concerns about CL robots being deployed in our daily lives, they mention that with further improvements CL robots could assist older adults and people with disabilities in their homes.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Online learning settings*; Supervised learning by classification;

Additional Key Words and Phrases: Continual learning, perceptions of robots, robot learning from human teachers, long-term human–robot interaction

This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

Authors' Contact Information: Ali Ayub (Corresponding author), University of Waterloo, Waterloo, ON, Canada; e-mail: a9ayub@uwaterloo.ca and Concordia University, Montreal, QC, Canada; e-mail: ali.ayub@concordia.ca; Zachary De Francesco, University of Waterloo, Waterloo, ON, Canada; e-mail: ; Jainish Mehta, University of Waterloo, Waterloo, ON, Canada; e-mail: jm3mehta@uwaterloo.ca; Khaled Yaakoub Agha, University of Waterloo, Waterloo, ON, Canada; e-mail: khaled.yaakoubagha@uwaterloo.ca; Patrick Holthaus, University of Hertfordshire, Hertfordshire, UK; e-mail: p.holthaus@herts.ac.uk; Chrystopher L. Nehaniv, University of Waterloo, Waterloo, ON, Canada; e-mail: cnehaniv@uwaterloo.ca; Kerstin Dautenhahn, University of Waterloo, Waterloo, ON, Canada; e-mail: kdautenh@uwaterloo.ca.

ACM 2573-9522/2024/10-ART52
https://doi.org/10.1145/3659110

**ACM Reference format:**
Ali Ayub, Zachary De Francesco, Jainish Mehta, Khaled Yaakoub Agha, Patrick Holthaus, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2024. A Human-Centered View of Continual Learning: Understanding Interactions, Teaching Patterns, and Perceptions of Human Users Toward a Continual Learning Robot in Repeated Interactions. *ACM Trans. Hum.-Robot Interact.* 13, 4, Article 52 (October 2024), 39 pages.
https://doi.org/10.1145/3659110

## 1 Introduction

We envision a future of general purpose assistive robots that can help users with a variety of tasks in dynamic environments, such as homes, offices, shopping malls, and so on [37, 43, 47, 51, 55]. It would be necessary that such assistive robots are personalized to their users' needs and their environments [5, 30, 52, 55]. However, over the long term, users' needs, preferences, and environments will continue to change, which makes it impossible to pre-program the robot with all the tasks it might be required to perform. A solution to this problem is to allow people to continually teach their robots new tasks and changes in their environments on the fly, an approach known as **continual learning (CL)** [8, 9].

CL has been extensively studied in recent years to allow robots to learn over long periods of time [8, 24, 45, 49, 58]. As it is imperative for a robot to learn and understand the objects in its environment [62, 64], the majority of research in CL has focused on **machine learning (ML)** models for object recognition in recent years [6, 24, 29, 46, 69]. Most of these techniques were tested on static object recognition datasets with a large number of training images for each object class. In real-world environments, however, robots will need to learn from individual interactions with their users who might be unwilling to provide a large number of training examples for each object.

In the past few years, robotics researchers developed CL techniques that can learn from only a few training examples per object, an approach known as **few-shot class incremental learning (FSCIL)** [3, 8, 60, 70]. Although FSCIL techniques produced promising results on object data collected from real robots, they were only tested with systematically collected "non-social" datasets by their experimenters [8, 9, 22]. Overall, most research in CL has been *robot-centered* or *algorithm-focused*, i.e., developing efficient CL algorithms that can learn from static datasets. However, in the real world, robots will learn from real users who might be unfamiliar with robot programming and learning. Therefore, an equally important area of research in CL is *human-centered*, to understand how human users interact with, teach, and perceive CL robots over the long term. Although human-centered AI has been argued to be a vital component for intelligent and autonomous robots [21], this concept is yet to be investigated within the domain of CL robots. To the best of our knowledge, we know of no other work involving long-term user studies where users teach modern CL models deployed on robots in real time over multiple interactions.

In this article, we have a human-centered focus to uncover the diversity and evolution of human teaching when interacting with a CL robot over repeated sessions. We developed a system for **socially guided CL (SGCL)** that integrates a **graphical user interface (GUI)** with CL models of object recognition deployed on the Fetch mobile manipulator robot [67]. We conducted a long-term between-participant study ($N = 60$) where participants interacted with and taught everyday household objects to a mobile manipulator robot (Fetch) that used three different CL models. We conducted 300 interactive sessions with 60 participants, where each participant taught a total of 25 objects to the CL robot in 5 sessions, with 5 objects per session. We analyzed the data collected in the study to characterize various aspects of human teaching of a CL robot in an unconstrained manner. Our results highlight the variation in the teaching styles of different users, as well as the influence of the robot's performance on users' teaching styles over multiple sessions. Our results indicate that the constrained experimental setups traditionally used to test most CL models are inadequate, as real users teach CL robots in a variety of ways. Finally, our results show that although users have concerns regarding the CL robot, most users can envisage personalized CL robots in the future, in particular for assisting older adults and people with disabilities in their homes.

A preliminary version of this work was presented in [4]. This article substantially extends [4] in several ways. First, to get a deeper understanding of the teaching style of the users for different CL models, we have reported results for a third condition with the **joint train (JT)** CL model. A total of 20 more participants were recruited in this condition who interacted with the CL robot with the JT model for 100 sessions. We redid all the experimental and statistical evaluations in Section 5 to incorporate all three conditions, which provided further insights into the interactions and teaching styles of the participants. We report detailed findings of the study in the updated discussions section (Section 6). Second, we report an analysis of the open-ended questions asked during audio interviews with each of the participants at the end of the study. Specifically, we transcribed all the audio files and performed semantic, thematic, and word-cloud analyses on the transcribed data. These analyses showed participants' perceptions toward CL robots deployed in everyday environments and suggestions for future improvement of these robots. A detailed discussion of the audio data analyses is reported in Section 6. Finally, the article contains a more complete background and related work discussion and provides a more detailed explanation of the methodology.

## 2 Related Work

In this section, we first present an overview of modern CL methods mostly tested without human users, and then introduce current approaches to robot teaching, highlighting the need for a human-centered approach at the intersection of CL and **human–robot interaction (HRI)**.

### 2.1 CL

The standard CL problem for an object recognition task is defined as: Suppose a CL model $\mathcal{M}$ gets a stream of labeled training datasets $D^1, D^2, \ldots$ over multiple increments, where $D^t = \{(x_i^t, y_i^t) : 1 \leq i \leq |D^t|\}$ is the dataset in the $t$th increment, $x_i^t$ is the $i$ the data point in $D^t$ with label $y_i^t$. $L^t$ is the set of object classes in the $t$th training dataset, where $L^j \cap L^k = \varnothing, \forall j \neq k$. During the testing phase, if the model $\mathcal{M}$ is given the increment label when predicting the class label of a data point, this setup is known as task-incremental learning [18, 36, 41]. In contrast, for the CIL setup, the model $\mathcal{M}$ is tested in increment $t$ on data points belonging to any of the previous classes $(L^1, \ldots, L^t)$ without access to the increment label [16, 32, 49, 69]. For example, consider that object classes $\{apple, orange\}$ are taught to the CL model in increment 1, and classes $\{banana, cup\}$ are taught in increment 2. In this case, 1 and 2 represent the increment labels, which are provided at test time in a task-incremental setup but not in CIL, i.e., in a task-incremental setup, at test time, the CL model will be provided with a test image and an increment label providing information about the increment in which the model was trained on the test object's class, whereas in a CIL setup, the model is only provided with a test image similar to the traditional classification setting. Although task-incremental setup can have some real-world applications, CIL is a more realistic CL setup for most human-in-the-loop robot learning applications, as robot users might not be willing to (or even remember) the increment label when asking the robot to predict the class label of an object. Therefore, we mainly review CIL works in this article.

*2.1.1 CIL.* One of the primary challenges faced by CIL techniques, and CL approaches in general, is catastrophic forgetting [23, 44]. Catastrophic forgetting[1] occurs when a CL model forgets previously learned knowledge when learning new information [49]. To address this issue, various research directions have been explored in the past to develop CIL models that can mitigate catastrophic forgetting [7, 16, 29, 49, 69]. One common strategy employed by existing CIL methods to avoid catastrophic forgetting is to store a portion of training samples from previous classes and then retrain the model on a mixture of the stored data and new data [16, 49, 69]. However, this approach faces scalability issues, especially when dealing with a large number of classes or long sequences of data, as storing and managing data from previous classes can quickly exhaust memory capacity, limiting performance in real-world applications. To address the memory storage problem, some CL approaches use regularization techniques [36, 41] without using any data from previously learned classes. Although these approaches alleviate the memory burden, their performance tends to be significantly inferior to methods that store old class data. Another set of approaches relies on generative replay to avoid storing raw data and instead generate old data using stored class statistics [6, 34, 46, 57, 68]. While generative methods offer memory efficiency, their performance can deteriorate drastically, especially when learning over longer sequences or dealing with complex datasets. In addition to the memory and performance concerns, one of the major issues faced by all CIL approaches is their performance degradation when dealing with limited training data [7, 60, 63]. This limitation makes them unsuitable for scenarios where the learning process must be based on a limited number of examples, such as when learning from human users who might be unwilling or unable to provide hundreds or thousands of images per object class.

*2.1.2 FSCIL.* In recent years, CL researchers have made significant progress in developing FSCIL models that can learn from just a few training examples per class [7, 60, 63]. In FSCIL, CL models are trained on a large number of base classes with a large dataset in the first increment, enabling them to learn a good general representation of the data. In subsequent increments, the model leverages the representation learned from the base classes to learn new classes with only a few training images per class [7, 12, 26, 59, 60, 70]. More formally, in an FSCIL setup, a CL model $\mathcal{M}$ receives a stream of labeled training datasets $D^1, D^2, ...$ over multiple increments, where $D^t = \{(x_i^t, y_i^t) : 1 \leq i \leq |D^t|\}$ is the dataset in the $t$th increment, $x_i^t$ is the $i$ the data point in $D^t$ with label $y_i^t$. $L^t$ is the set of object classes in the $t$th training dataset, where $L^j \cap L^k = \emptyset, \forall j \neq k$. $D^1$ is the large-scale training dataset of base classes to learn a good representation for the model, and $D^t, \forall t > 1$ is the few-shot training set of new classes. After training on $D^t$, $\mathcal{M}$ is tested to recognize all encountered classes in $L^1, ..., L^t$. For $D^t, \forall t > 1$, this setting with $C$ classes and a few $K$ training samples per class is known as the *$C$-way $K$-shot* FSCIL. Note that few-shot learning [19], unlike FSCIL, is not trained on a stream of *$C$-way $K$-shot* tasks, and the ML model is required to classify test images belonging to the new $C$ classes only, and not the base classes.

In prior research, FSCIL approaches were only tested on static, datasets captured in constrained setups (e.g., Modified National Institute of Standards and Technology [40]) and not on physical robots that might not have perfect data available. Although a few FSCIL approaches [8, 22, 64] have been developed for learning on real robots, none of these FSCIL approaches were tested with real participants. Instead, they were tested on datasets that were captured by the experimenters on robots in systematically controlled setups. Real users, however, might not be aware of the underlying CL models and might provide imperfect or inconsistent data during teaching or testing.

[1] Note that the term *catastrophic forgetting* is mainly used in ML literature to describe the phenomenon of an ML model forgetting most past knowledge. However, when interacting with real users, the perception of forgetting might be far from "catastrophic."

Understanding how human users perceive CL systems on robots is crucial to ensure that these systems are user-friendly, intuitive, and reliable.

## 2.2 Human–Robot Teaching

Human-centered research for robot learning through HRI has been relatively limited, with only a few user studies conducted in the past to explore the characteristics of human teaching with simulated and real robots. However, most of these studies were performed in Wizard of Oz setups, where the robot did not actually learn from human teaching [20, 33, 48]. While interactive reinforcement learning through HRI has been explored for learning manipulation tasks, kitchen-related tasks, and understanding natural language descriptions of images from humans [10, 14, 38, 56], most of these studies were primarily focused on testing the performance of reinforcement learning models for robot manipulation or investigating users' perceptions toward these models, rather than examining patterns of human teaching. Furthermore, these studies were limited to single interaction sessions with users, not fully capturing the dynamics and evolution of human teaching over time. Thomaz et al. [62] conducted a study on object learning where a robot learned object names and simple affordances from interactions with human participants. Human participants taught six simple objects, such as blocks and spheres, to a social robot, which used a support vector machine-based method for learning these objects. While this study showed significant performance differences when ML models learned from human teachers compared to systematically collected datasets, it was also conducted in a single interaction session with participants. For CL robots, understanding how human teaching evolves over the long term is crucial to develop more effective and adaptive learning systems. In contrast, to the best of our knowledge, we conducted the first long-term user study at the intersection of continual ML and HRI. The study aims to understand patterns of human teaching in the context of CL robots over multiple interactions. By conducting a multi-session study, we can gain valuable insights into how human teaching strategies evolve and adapt over time when teaching CL robots.

## 2.3 Research Questions and Hypotheses

Based on the unique ways people interact with and teach robots [31, 35, 61], we consider the following research questions (RQs):

*RQ1:* How do different users label objects when teaching a CL robot over multiple sessions?
*RQ2:* How does a CL robot's performance affect the way users teach over multiple sessions?
*RQ3:* Do users change the way they teach a CL robot over multiple sessions?
*RQ4:* Is there a difference in teaching style and robot performance for expert and non-expert users?
*RQ5:* How do human users perceive a CL robot for everyday applications?

We analyze the data collected in our study to answer the five research questions and test the associated hypotheses. The following hypotheses are guided by previous research that was discussed Sections 2.1, and 2.2: Prior HRI research showed that users' interactions and perceptions toward a robot are correlated with the performance of the robot and the time and effort spent in interacting with the robot. Also, there might be differences in how different users interact with the robot, especially if they had prior experience programming robots. Further, prior CL research showed that CL models can forget previous knowledge over time, and thus their performance decreases. However, there is a difference in the rate of forgetting for different CL models. Finally, prior HRI research showed that the novelty effect can wear off in long-term interaction studies, therefore users' excitement to interact with the robot would decrease.

Note, H*n*.*m* is the *m*th hypothesis related to the research question *n*, e.g., H1.3 is the third hypothesis to answer RQ1.

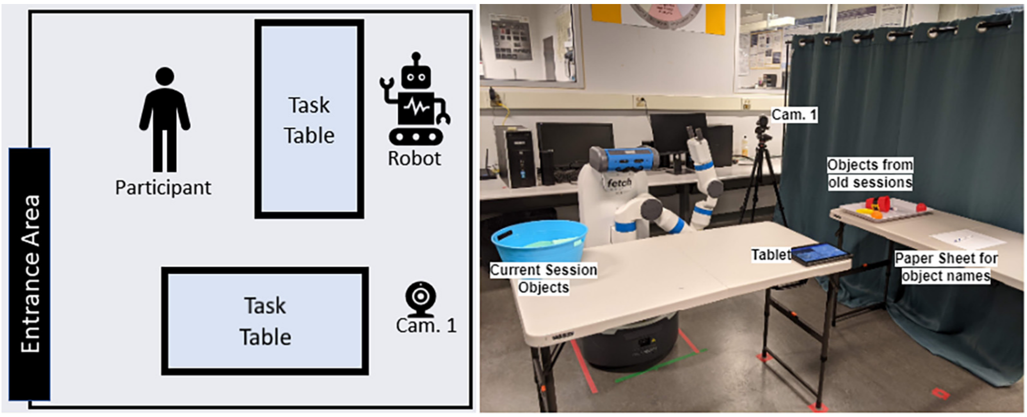


Fig. 1. (Left) Experimental layout for the SGCL setup with the participant and the robot. (Right) Corresponding real-world setup. Cam. 1, camera 1.

- *H1.1:* Labeling strategies for objects vary among different users.
- *H2.1:* Classification performance of the robot affects the teaching style of the participants over multiple sessions.
- *H2.2:* Users teach a robot that forgets previous objects differently than a robot that remembers previous objects.
- *H2.3:* Users teach a robot that retrains on all previous objects differently than a robot that does not train on all previous objects.
- *H3.1:* Teaching styles of users change over multiple sessions regardless of the CL model.
- *H4.1:* CL robots taught by expert users perform better than the ones taught by non-expert users.
- *H4.2:* There is a difference between the teaching styles of expert and non-expert users.

As RQ5 is related to the open-ended questions asked to the participants, we do not have formal hypotheses for this question. However, we expected that participants would perceive the robot to be useful for everyday applications, particularly for older adults and people with disabilities.

## 3 SGCL System

We investigated human teaching patterns when interacting with a CL robot to teach an object recognition task. In this experiment, in each session, the user taught the robot household objects in a table-top environment and then tested the robot to find and point to the requested object on the table. Figure 1 shows the table-top experimental setup for this study. The setup and the task are straightforward which makes it clear what the user should do to teach the robot different objects, and what the robot should do to find the learned objects during the testing phase.

For this setup, we developed a CL system for the object recognition task, which integrates CL models with a Fetch mobile manipulator robot [67], as well as a GUI for interactive and transparent learning from human users. Figure 2 shows our system for the object recognition task. In this system, the user interacts with the robot through the GUI on an Android tablet (Figure 3). Formally, in each session (or increment) $t$, the user teaches the robot $L_t$ number of objects by placing them in front of the robot and labeling them using the GUI. The images of $L_t$ objects are captured by the robot and then pre-processed using the object detection module. Combined with the labels of the images from the user, a dataset $D^t = \{(x_i^t, y_i^t) : 1 \leq i \leq |D^t|\}$ is generated, where $x_i^t$ is the $i$th image in the dataset with the class label $y_i^t$. The CL model $\mathcal{M}$ is then trained on the dataset

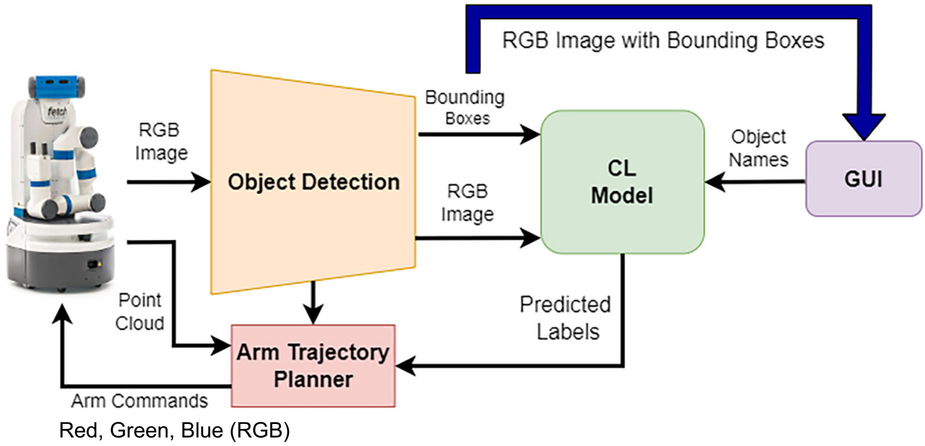


Fig. 2. Our complete SGCL system. Processed RGB images from Fetch's camera are sent to the GUI for transparency and also passed on to the CL model. The user sends object names to the CL model either for training the CL model or finding an object. The arm trajectory planner takes point cloud data, processed RGB data, and predicted object labels from the CL model as input and sends the arm trajectory for the Fetch robot to point to the object.

$D^t$. Note that unlike traditional CL setups (such as, CIL and FSCIL described in Sections 2.1), the number of images per object class in a session is not fixed as it is dependent on the number of times the user teaches an object to the robot. Further, there can be an overlap in the object classes taught in different sessions depending on how the user labels the objects. For example, the user can name two different cups with different names, such as "green cup" and "red cup," or they can name both of the cups as "cup."

After teaching, the user can test the robot by asking it to find objects on the table through the GUI. The robot passes the pre-processed images to the CL model to get the predicted object labels. If the object requested by the user is found, the robot finds the **three-dimensional (3D)** location of the object on the table and points to the object using its arm. Note that the user has flexibility in terms of the total number of objects to be tested in an increment, as well as which objects to test (old or new objects). Due to this flexibility in SGCL, results for CL models were quite different from the results on static datasets (see Section 5 for details). Our code for the complete systems is available at https://github.com/aliayub7/cl_hri.

## 3.1 CL Models

The goal of our study is to do an in-depth analysis of how users interact with and teach CL models over long-term interactions, with one of the questions being if the performance of the model has an effect on the teaching style of the users. For such an analysis, it is imperative to choose a meaningful baseline. The naïve **fine-tuning (FT)** approach [7, 49] has been used as the baseline on static "non-social" datasets in the CL literature. Therefore, we chose to test FT as our study's baseline model. In this approach, a **convolutional neural network (CNN)** [25] is trained on the image data of the object classes in each increment (i.e., in an interactive session with the user). The model does not train on any of the objects learned in the previous increments (sessions), and therefore, it catastrophically forgets the previously learned objects. More details about this model can be found in [7, 49, 69].

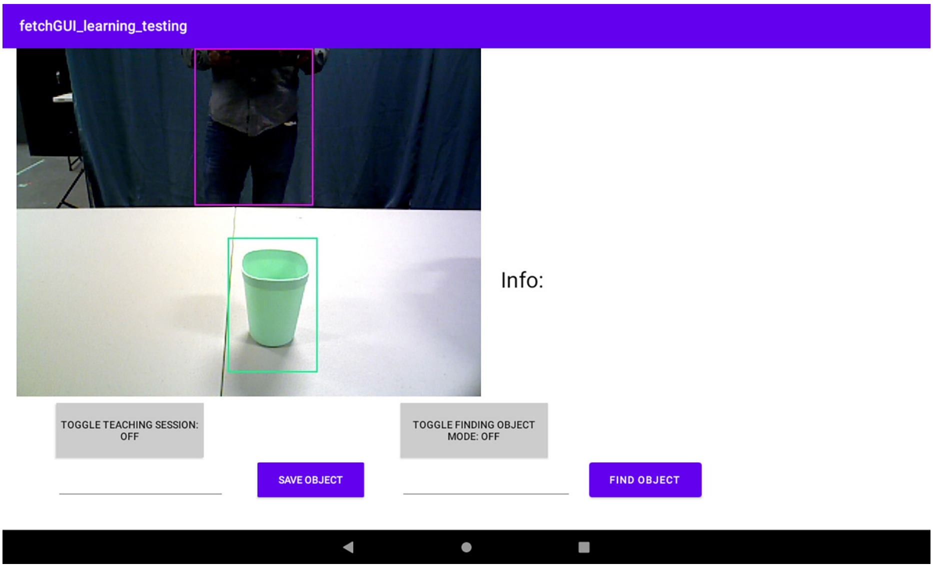

fetchGUI_learning_testing
Info:
TOGGLE TEACHING SESSION: OFF
TOGGLE FINDING OBJECT MODE: OFF
SAVE OBJECT
FIND OBJECT

Fig. 3. The GUI used to interact with the robot. The RGB camera output with bounding boxes is on the top left. The buttons at the bottom can be used to teach objects to the robot and ask it to find objects in the testing phase. The top right of the GUI shows information sent by the robot to the user.

For the second model, we consider a state-of-the-art CL approach specifically designed for FSCIL in robotics applications [8]. This approach termed **centroid-based concept learning (CBCL)** uses a CNN pre-trained on the ImageNet dataset [53] to extract feature vectors for the images of objects in each increment $t$. CBCL then clusters the feature vectors of all the object classes in the increment $t$ and generates a set of centroids $C^y = \{c_1^y, ..., c_{n_y}^y\}$ for each object class separately, where $n_y$ is the total number of centroids for class $y$. CBCL mitigates forgetting by generating separate clusters for different classes in increment $t$.

For the classification of a new object, CBCL finds the distance of the feature vector of the test object from the centroids of the learned object classes. CBCL then uses a weighted voting scheme to find the most common class among the closest centroids to the test feature vector. The most common class is predicted as the object class for the test feature vector. More details about CBCL can be found in [7].

Finally, for the third model, we consider the batch learning (called JT in this article) approach that stores the image data of the object classes from previous increments and retrains using stored data when learning new objects. JT has been used in CL literature as a theoretical upper bound for CL on static datasets. JT trains a CNN model on a combined image dataset of the new and old object classes in each increment, and therefore, its training time continues to increase with each increment. Note that the only difference between JT and FT is that FT trains only on the objects taught in an increment, whereas JT trains on all the objects taught so far in each increment. More details about JT can be found in [16, 49].

Note that all of these models were only tested on systematically collected object datasets by experts in prior research [8, 42, 49] and have never been trained or tested in real-time with human participants. In this article, we integrate FT, CBCL, and JT in a fully autonomous system that allows users to experience these different ML models in real time through the Fetch mobile manipulator robot [67].

## 4 Method

### 4.1 Artifacts—Fetch Mobile Manipulator Robot

As manipulator robots with RGB-D cameras are suited for the recognition and manipulation of objects, we use the Fetch mobile manipulator robot [67] in our experimental setup (Figure 1). The Fetch robot is equipped with a 7 df arm, a mobile base, an RGB camera, a depth sensor, and a Lidar sensor. The robot's sensors play a crucial role in enabling 3D perception, simultaneous localization and mapping, and obstacle detection in the world. In our specific setup, we do not require the robot to manipulate objects or move its base but ask it to point to requested objects using its arm. This intentional design choice allows us to focus solely on CL, which primarily involves the process of learning and recognizing objects. To ensure the safety of human participants in our study, we conducted a thorough safety analysis that was approved by the University of Waterloo's Human Research Ethics board. In addition, we implemented various mitigating strategies to minimize potential risks. For example, before moving its arm, the robot always reminded the participant to stay at a safe distance. Further, the experimenter was observing the entire interaction and could disable the robot remotely using an emergency stop button if there was a risk of collision with the participant. As a result, the robot was deemed safe to be used in conjunction with human participants during our research study.

To handle the possibility of multiple objects present on the table within the robot's camera view, we employ additional processing steps for the RGB images. These images are passed through a generic object detector [50] to identify regions within the image that are likely to contain objects (refer to Figure 3). To refine the detected regions and eliminate any overlapping detections, we apply non-max suppression [28], resulting in an accurate representation of distinct objects. Additionally, we leverage the depth perception of objects to filter out any detected objects that do not reside on the table. This helps exclude background objects and any instances where a participant may be interacting with the robot (as illustrated in Figure 3). The resulting regions of interest, representing objects detected on the table, are cropped into separate images and then forwarded to the CL model.

### 4.2 GUI

To facilitate interactive and open-ended teaching of different objects for the users, we developed a GUI deployed on an Android tablet. A screenshot of the GUI is presented in Figure 3. The GUI's top left section displays the pre-processed camera output from the robot, featuring bounding boxes that indicate detected objects. This visual representation serves as a transparency device so that the participants can understand what the robot perceives on the table. Located at the bottom left of the GUI is a toggle button that can be pressed to initiate a teaching session with the robot. When pressed, the button turns green, signifying that the system has entered the teaching phase. During this phase, participants can input the name (class label) of the objects in the space below the toggle button and save an image of the object by using the save button adjacent to the input area. On the bottom right of the GUI, another toggle button is situated, enabling participants to commence the testing phase. Similar to the teaching button, this button turns green once pressed. In the testing phase, participants can enter the name (class label) of the object they want the robot to locate on the table in the space below the toggle button. By pressing the "Find Object" button next to the input area, participants can instruct the robot to search for the specified object on the table. Lastly, the top right section of the GUI displays messages communicated by the robot to the user during the session. These messages are also vocalized using a text-to-speech module available in robot operating system.

We intentionally opted not to utilize a **natural language processing (NLP)**-based interaction system for teaching objects due to the inherent challenges associated with designing an open-ended

NLP system, as highlighted in [17]. Such systems involve speech-to-text and natural language understanding phases, which introduce potential errors and could divert participants' attention from the primary focus of our research study: the CL of household objects. Our objective was to investigate participants' interactions, teaching styles, and perceptions of the CL system rather than evaluate the robot's communicative capabilities. By avoiding the complexities of an NLP-based system, we aimed to maintain a streamlined research environment, improve robustness, and minimize potential distractions from the main research goal.

### 4.3 Participants

We recruited 63 participants (35 **female (F)**; 28 **male (M)**, all students) from the University of Waterloo, between the ages of 18 and 37 years ($\mu = 23.12$, $\sigma = 4.04$). Three out of 63 total participants did not complete the five sessions for the study; therefore, we report results for 60 participants who completed the study. All three conditions were randomly assigned 20 participants each (ages: $\mu = 22.7$, $\sigma = 4.58$, 9 F, 11 M for *CBCL* condition, ages: $\mu = 24.53$, $\sigma = 4.03$, 10 F, 10 M for *FT* condition, ages: $\mu = 22.2$, $\sigma = 2.96$, 15 F, 5 M for *JT* condition). Analysis of the pre-experiment self-assessment survey showed that 35% of the participants were familiar with robot programming, 55% reported that they had previously interacted with a robot, 5% were familiar with the Fetch mobile manipulator robot, and 8% had previously participated in an HRI study. For the rest of the article, we call participants with prior robot programming experience "experts" and the rest of the participants "non-experts." Experts and non-experts were randomly assigned to each of the three conditions. A Chi-square test between the previous programming experience of the participants and the assigned condition ($\mathcal{X}^2(1) = 2.45$, $p = 0.29$) confirmed that there is no statistically significant correlation between these two variables; and therefore, the distribution is not significantly different from uniform. All procedures were approved by the University of Waterloo Human Research Ethics Board.

### 4.4 Procedure

We conducted five repeat sessions with each participant in a robotics laboratory on campus, with each session lasting approximately 20–30 minutes. All sessions were video-recorded for further analysis. Participants were randomly assigned to one of the three experimental conditions, each utilizing a different CL model: CBCL, FT, or JT. Before their first session, participants were asked to complete an online consent form and a pre-experiment survey using the Qualtrics platform [1]. Upon completing the survey and the consent form, the experimenter provided a brief oral introduction to the experiment and greeted the participant. The participant was told to consider the Fetch robot as their personal household robot, with the opportunity to teach the robot various household objects over a period of 5 days, gradually accumulating a repertoire of 25 objects. Subsequently, the participant was directed to the designated study area within the laboratory and was given an Android tablet equipped with the GUI shown in Figure 3. They were informed that initially, they would run a demo session with a robot, allowing them to familiarize themselves with the process of teaching and testing the robot. The participant was made aware that the robot would not be learning any of the objects presented in front of the camera during this demonstration phase.

The experimenter provided a detailed explanation of the session structure to the participant, outlining that each session with the robot would consist of two phases: a teaching phase and a testing (i.e., finding an object) phase. To illustrate this, the experimenter utilized a blue cup as a demonstration object (which would not be used later in the study) and placed it on the table. The participant was then directed to stand in the designated area facing the table and instructed to initiate a teaching session by pressing the "Toggle Teaching Session" button on the GUI. Once pressed, the button turned green, accompanied by a message displayed on the tablet from the robot stating, “Entered teaching mode. You can now start teaching me objects.” The robot vocalized the same message through its speakers. Next, the experimenter asked the participant to type the name of the object into the text box located below the toggle button. The participant was informed that they had the freedom to choose any name for the object. After naming the object, the participant was instructed to save it by pressing the “Save Object” button situated next to the text box. Upon pressing the button, the robot stated, “[OBJECT NAME TYPED IN THE TEXT BOX] has been saved.” The experimenter also informed the participant that they could save each object multiple times by placing it at different locations on the table and varying its angles. They further mentioned that a similar procedure can be used to teach the other four objects in the session. The participant was further informed that once they had saved all five objects in a session, they could end the teaching phase by pressing the toggle button again. When deactivated, the button turned grey, and the robot stated, “I am learning the objects, please wait.” Subsequently, the robot stated, “Left teaching mode,” indicating that it had completed the learning process and exited the teaching mode. The experimenter clarified to the participant that the robot would now process and learn the objects demonstrated during the session and would later communicate when it had finished learning and exited the teaching mode.

The experimenter proceeded to explain the testing phase to the participant. They instructed the participant to initiate the testing phase of the session by pressing the “Toggle Finding Object Mode” button on the GUI. Once pressed, the button turned green, accompanied by the robot’s statement, “Entered finding mode.” The experimenter then placed two additional objects on the table alongside the demo object, resulting in a total of three objects on the table. Participants were informed that during the testing phase, they had the option to place one or up to three (a suggestion, not a requirement) objects on the table. The experimenter mentioned that after placing the objects on the table, the participant could enter the name of the object they wanted the robot to find in the text box located below the toggle button. Specifically, the experimenter requested the participant to type “cup” in the text box to instruct the robot to find that particular object on the table. Once the participant finished typing, they were instructed to press the “Find Object” button. Upon pressing the button, the robot responded, “I will point to the cup now. Please make sure that you are at a safe distance from me.” Subsequently, the robot moved its torso and arm to point at the cup on the table. Upon completion, the robot stated, “I am done.” The experimenter highlighted that the participant could ask the robot to find objects taught in the current session and previous sessions by placing them on the table and utilizing the “Finding Object Mode.” The experimenter then requested the participant to press the “Toggle Finding Object Mode” button once again to exit the testing phase. When deactivated, the button turned grey, and the robot announced, “Left finding mode.” Note that during the demo phase, the robot was able to find the object every time, however, when participants tested the robot on their own, the robot did make errors and could not find objects. In those cases, the robot stated, “I cannot find [OBJECT NAME TYPED IN THE TEXT BOX],” and did not move its arm.

After the completion of the demonstration phase, lasting ~5 minutes, the experimenter provided a paper sheet to the participant, serving as a memory aid. The participant was instructed to write down the names of the objects taught in the current session on the sheet. The experimenter collected and retained the paper sheet, returning it to the participant at the beginning of each subsequent session. This practice allowed participants to recall the object names when needed for instructing the robot to find those objects in subsequent sessions.[2] Subsequently, the experimenter retrieved

[2] It is important to note that in real-world scenarios, participants would not need to rely on a written sheet, as they would have ongoing interactions with the objects. However, in our setup, due to the limited duration and time gaps between sessions, it was necessary to provide a means for participants to remember the object names. On average, the time between the first and last session was 25.3 days, with a maximum of 52 days. Additionally, the average duration between any two consecutive sessions was 6.2 days, with a maximum of 29 days.

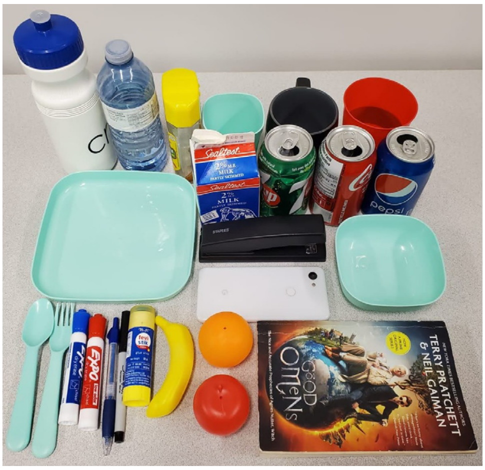


Fig. 4. The 25 objects used in our study. Note that the visual similarity of some objects and their size variation make this a challenging task.

the tablet from the participant and loaded the program for the actual session on the tablet. Once prepared, the experimenter returned the tablet to the participant. Simultaneously, the experimenter placed five objects to be taught in the session on one side of the table. At this point, the experimenter informed the participant that they could commence their session and begin the process of teaching the five objects.

The experimenter then moved to a secluded area, and the participant began teaching the robot the five designated objects. Once the teaching phase was completed, the participant transitioned to the testing phase, during which they instructed the robot to find objects from both the current session and previous sessions. Note that the teaching and testing phases were not fixed for all participants, because the participants had flexibility in terms of how they wanted to teach the robot. For example, participants could reteach any objects that the robot misclassified during the testing phase, and this could be repeated as many times as they desired. Once the testing phase concluded, the experimenter returned from the secluded area and expressed gratitude, saying, "Thank you for coming today. We have a few questions about your experience today. Could you please answer them on this tablet?" To facilitate this, the experimenter provided a separate tablet to the participant, containing questionnaires in the Qualtrics [1] format. Results from the post-experiment questionnaire data are reported in a separate work [2]. Following the completion of the questionnaires, the experimenter thanked the participant for their participation. Subsequently, the participant scheduled their next session.

In the subsequent four sessions, which lasted ~20–30 minutes each, the same procedure was followed with a few variations. The objects taught during each session were different, and a total of 25 objects were used throughout the five sessions. Figure 4 shows the 25 objects utilized in our study. During sessions 3–5, participants were informed that they had the option to bring up to two objects of their choice to teach the robot. In such cases, some objects from our predefined set (Figure 4) were replaced with the participants' objects. However, the total number of objects taught remained consistent at 25 over the five sessions. Further, participants did not engage in a demo interaction in the subsequent four sessions. At the conclusion of the final session, participants were requested to participate in a brief interview aimed at capturing their experience with the robot. These interviews were audio-recorded for further analysis. Participants were compensated with

$30 CAD if they completed all five sessions. Alternatively, if participants did not complete all five sessions, they received $6 CAD per session completed.

### 4.5 Measures

We used both qualitative and quantitative measures to analyze the data for the three conditions. We analyzed the object names given by the participants to different objects using the image data stored for objects during teaching sessions. We report the variety and frequency of labels used by the participants for each object. This measure was used to help answer RQ1 about the labeling styles of different users. We also coded the video recordings to calculate the frequency of teaching by the participants in all five sessions, and if they re-taught any objects to the robot in case the robot was not able to correctly find them on the table, to help answer RQ2 about the effect of the robot's performance on users' teaching styles.

We also analyzed the performance of three CL approaches. Classification accuracy per session (increment) has been commonly used in the CL literature [49, 60] for quantifying the performance of CL models for object recognition tasks. Therefore, for each session, during the testing phase, we recorded the total number of objects tested by the participant and the total number of objects that were correctly found by the robot. Using this data, we calculated the accuracy $\mathcal{A}$ of the robot in each session as

$$\mathcal{A} = \frac{number\ of\ objects\ correctly\ found\ in\ the\ session}{number\ of\ objects\ tested\ in\ the\ session}. \tag{1}$$

We use the accuracy of the models to determine the teaching quality of the participants in each condition and over multiple sessions. This measure was used to help answer RQ2 related to the performance of the robot and its effect on the teaching styles of the users. Further, using the image data stored for the objects, we calculated the average number of times each object was taught by the participants in each session to determine the effort spent by the participants in teaching the robot. We also analyze how the above-mentioned variables are affected by the sessions, choice of the CL model, and previous robot programming experience of the participants, to help answer RQs 3 and 4 that are related to the effect of session and previous programming experience of the participants on their teaching styles.

For the open-ended interviews, one of the questions is quantitative asking participants about rating the teaching and testing process ranging from 1 (boring) to 5 (exciting). For the remainder of the questions, we perform aspect-based sentiment analysis on the data collected from the open-ended audio interviews to understand user perceptions toward the CL robot. We use the Python **natural language toolkit (NLTK)** [13] for sentiment analysis for each open-ended question (Table 1) across 60 participants. The output of the NLTK sentiment analysis provides a compound variable with values ranging from −1 to 1, where values closer to 1 show highly positive sentiment, values closer to 0 show neutral sentiment, and values closer to −1 show highly negative sentiment. Finally, we analyze participants' responses using the six-step **thematic analysis (TA)** method developed by Braun and Clarke [15]. We first generate category codes after analyzing the data, and then determine themes in the data where a theme is accepted if we find the corresponding category codes at least 30 times across three conditions. Themes and category codes are then reviewed and revised. Two of the authors independently conducted the TA on the dataset. We also perform word cloud analysis of the open-ended responses to visualize the most common words in the responses. These measures were used to help answer RQ5 about user perceptions of the robot for everyday applications.

Table 1. List of Open-Ended Questions asked to the Participants

| No. | Keyword | Questions |
|---|---|---|
| 1 | robot_perception | What do you think about this robot or system? |
| 2 | robot_house | How do you feel about this robot in your house? |
| 3 | assist_friend | Would you consider the robot for a relative or friend? |
| 4 | concern_robot | What concerns would you have about this robot assisting with daily tasks? |
| 5 | suggestion_improvement | What things should robot developers focus on to make this robot to be used in households? |
| 6 | pre-defined_objects | Would it be better for the robot to come with a pre-defined set of objects or should the users be able to directly teach the objects in their homes? |

Keywords are added to refer to the questions in a meaningful way.

## 5 Results

In this section, we present the results of our analysis in terms of different labeling strategies and teaching styles of the participants. We also report the effect of participants' teaching styles on the robot's performance, and vice versa.

### 5.1 Object Labeling by Human Teachers

Table 2 shows the number of different labels given to the 25 objects by 60 participants in the study. To identify each object we add a generic name for each object in the table. For example, for the plastic apple used in our study, we identify it as an apple in the table. Overall, there was a significant variation in the labeling of objects by the participants, ranging from 5 (for Banana) to 18 (for Honey) different labels for objects. Among such labels, some were quite simple and generic, such as *Honey*, *Bowl*, *Milk*, and so on whereas some were quite specific, such as *Almost Empty Yellow Honey Jar*, *Light Green Flat Bowl*, *Empty Milk Carton*, and so on. We also report the most common label given to each object and the percentage of participants that chose that label. The consensus among the participants for labeling the objects varied from 30% for Red Cup to 90% for Banana. Overall, only 6 out of 25 objects had a label with high consensus among the participants (≥80% participants).

We also noticed some unique labeling strategies by the participants. Some of the participants labeled different objects in different sessions using the same label. For example, multiple participants gave the label *Cup* to *Green Cup* in Session 1, *Red Cup* in Session 2, and *Mug* in Session 3. In total, 16 out of 60 participants (26.7%) gave the same label to at least two different objects. Further, some participants gave multiple labels to the same objects. For example, one participant labeled *Milk* as both *Milk Box* and *Milk Pouch*. Overall, there were 13 out of 60 participants (21.7%) that gave more than one label to at least one object.

We also noticed that some participants gave unexpected labels to the objects. One participant gave Chinese labels to objects, such as *Niunai* for Milk, *Yuanzhubi* for Pen, and so on. Another participant gave German labels to some objects, such as *Kugelschreiber* for Pen, *Buch* for Book, *Tacker* for Stapler, and so on. A few participants also gave labels that did not match the objects. For example, one participant labeled Red Cup as *Mary*, Glue as *Tom*, Apple as *Mars*, and so on. In such cases, we noticed that some participants forgot in later sessions how they labeled different objects. Therefore, they asked for the robot's help to identify the object associated with the labels. For example, in their session 2, one participant asked the robot to find "Tom," and the robot pointed at Glue, thus helping the participant recognize that they had named Glue "Tom."

Table 2. The Number of Different Labels Given by the Participants to All 25 Objects in the Study Together with the Most Common Label for Each Object with the Percentage of Participants That Chose This Label

| Object | Number of Different Labels | Most Common Label | Examples |
|---|---|---|---|
| Green cup | 11 | Cup (67%) | Green cup, plastic cup |
| Honey | 18 | Honey (50%) | Container, bottle of honey |
| Bowl | 12 | Bowl (67%) | Green bowl, tiny bowl |
| Glue | 8 | Glue (75%) | Tom, glue stick |
| Spoon | 8 | Spoon (85%) | Aquamarine spoon, plastic utensil |
| Apple | 7 | Apple (85%) | Toy apple, red apple |
| Banana | 5 | Banana (90%) | Plastic banana, ban |
| Red cup | 15 | Red cup (30%) | Red, santa cup |
| Blue marker | 11 | Marker (62%) | Text marker, expo |
| Orange | 7 | Orange (83%) | Org, toyo orange |
| Mug | 8 | Mug (74%) | Makbei, tea cup |
| Fork | 7 | Fork (75%) | Aquamarine fork, teal fork |
| Sharpie | 9 | Sharpie (49%) | Marker, pen, black sharpie |
| Plate | 12 | Plate (70%) | Green tray, big plate |
| Stapler | 7 | Stapler (88%) | Punch, pin |
| Book | 5 | Book (89%) | Novel, interesting book |
| Red marker | 6 | Red marker (45%) | Red pen, dry erase, expo marker |
| Blue pen | 10 | Pen (57%) | Ball pen, Kugelschreiber, Yuanzhubi |
| Pepsi | 9 | Pepsi (56%) | Pepsi can, soda, Kele |
| White bottle | 11 | Water bottle (50%) | Bottle, Shuibei, sipper |
| Coca cola | 8 | Coke (41%) | Coke can, cola |
| Milk | 8 | Milk (73%) | Milk box, cow juice |
| Phone | 5 | Phone (76%) | Cell, mobile |
| 7Up | 15 | 7Up (43%) | 7Up can, soft drink |
| Water bottle | 8 | Water (46%) | Water bottle, plastic bottle |

Examples of labels given by the participants for each object are shown in the last column. Objects are ordered from top to bottom as they were taught in five sessions with five objects per session. Note that the first column shows some reference names for the objects to be able to identify them individually in the paper.

### 5.2 Participants' Teaching Styles and Robot Performance

We performed a three-way ANOVA with three independent variables: the three conditions (CBCL, FT, and JT), session number, and previous robot programming experience of the participants. The ANOVA was performed to understand the effect of the three independent variables on the teaching style of the participants and the robot's performance in the testing phase. The dependent variables were the following:

- *Classification accuracy:* The accuracy of the robot to find correct objects in the testing phases of 300 sessions with 60 participants (calculated using Equation (1)).

Table 3. Results (*p* Values and Effect Sizes) of the Three-Way ANOVA Using Session Number, Continual Learning Model, and Previous Programming Experience as Independent Variables

| | *p Values* | | | | |
|---|---|---|---|---|---|
| | **Accuracy** | **Number of Images** | **Teaching Phases** | **Reteaching** | **Old Objects** |
| Programming experience | 0.5772 | 0.0845 | 0.4399 | 0.8533 | 0.5517 |
| Session number | <**0.0001** | **0.0257** | 0.3903 | 0.0939 | **0.0153** |
| CL model | <**0.0001** | 0.3748 | 0.8394 | 0.3097 | 0.4634 |
| Programming experience: Session number | 0.5607 | 0.0534 | 0.5083 | 0.4564 | 0.2803 |
| Programming experience: CL model | 0.856 | 0.0868 | 0.6896 | 0.669 | 0.1634 |
| CL model: Session number | **0.0011** | 0.2624 | 0.6151 | 0.4653 | 0.1185 |
| Programming experience: CL model: Session number | 0.9111 | 0.0993 | 0.3127 | 0.2818 | **0.0256** |
| | *Effect Sizes* | | | | |
| | **Accuracy** | **Number of Images** | **Teaching Phases** | **Reteaching** | **Old Objects** |
| Programming experience | 0.0027 | 0.0424 | 0.0091 | 0.0003 | 0.0027 |
| Session number | 0.0714 | 0.008 | 0.0034 | 0.0199 | 0.0403 |
| CL model | 0.2467 | 0.0267 | 0.0053 | 0.0211 | 0.0117 |
| Programming experience: Session number | 0.007 | 0.0065 | 0.0026 | 0.0083 | 0.0139 |
| Programming experience: CL model | 0.0027 | 0.0702 | 0.0112 | 0.0072 | 0.0276 |
| CL model: Session number | 0.0674 | 0.0063 | 0.005 | 0.0177 | 0.0375 |
| Programming experience: CL model: Session number | 0.0074 | 0.009 | 0.0079 | 0.0232 | 0.0542 |

Columns for the accuracy of the models, number of images per object, number of teaching phases, reteaching misclassified objects, and old objects show *p* values (top table) and effect sizes (bottom table) for the dependent variables. We use generalized eta [39] to calculate effect sizes. Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are in bold.

- *Number of images per object:* The average number of images per object shown by the participants to teach the robot in each session.
- *Number of teaching phases:* The average number of times participants started a teaching phase in each session. Participants were not confined to teaching objects in a single teaching phase, and they could start multiple teaching phases in a session to teach objects to the robot.
- *Reteaching misclassified objects:* The average number of times participants retaught misclassified objects in each session. If an object was misclassified in the testing phase, it could be retaught to the robot by the participants.

Table 3 represents the *p* values and significance levels for the ANOVA. For classification accuracy, we see a significant effect based on the session number and the choice of the CL model (CBCL, FT, or JT condition) and the interaction between the session number and the CL model. For the number of images taught per object, we noticed a significant effect based on the previous programming experience of the participants and the interaction between the CL model and the programming experience. For the number of teaching phases per session, we noticed a significant effect based on the previous programming experience of the participants. Finally, for reteaching misclassified objects, we saw a significant effect of the choice of the CL model.

We noticed that the data for sub-groups for some dependent variables were not normally distributed. Although ANOVAs are fairly robust to outliers, we still performed a follow-up ANOVA with transformed values using the Box-Cox transformation [54]. Results for the transformed values were mostly consistent with the ones for the original values, with some differences. *Post hoc* Tukey's **honestly significant difference (HSD)** tests on significant ANOVAs for transformed values were consistent with significant ANOVAs for original values.

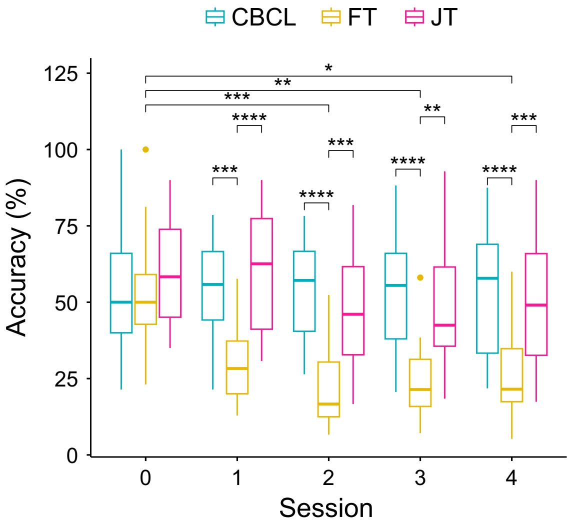
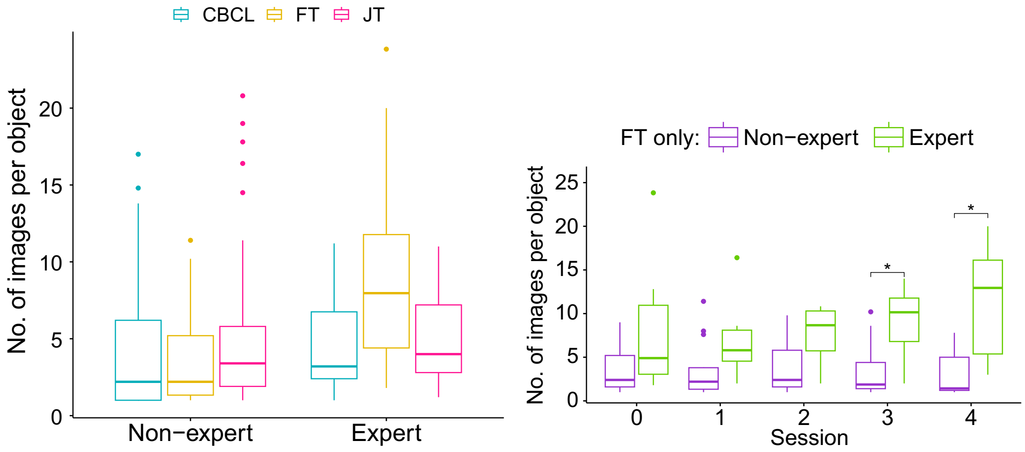


Fig. 5. Boxplot for *accuracy* for three conditions with different CL methods (CBCL, FT, and JT). Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are indicated on bars between columns.

*5.2.1 Model Accuracy.* We performed the *post hoc* Tukey's HSD test for the significant ANOVAs for model classification accuracy as the dependent variable. Figure 5 shows the average classification accuracy of the CL robot over five sessions. The accuracy is significantly affected by the choice of the CL model. For the first session, all three models have similar accuracy ($\mu = 0.53$, $\sigma = 0.19$ for CBCL; $\mu = 0.52$, $\sigma = 0.18$ for FT; $\mu = 0.59$, $\sigma = 0.18$ for JT). For the next four sessions, there is a statistically significant difference between FT and CBCL, and between FT and JT: when comparing FT ($\mu = 0.29$, $\sigma = 0.13$) to CBCL ($\mu = 0.54$, $\sigma = 0.16$) with $p = 0.0004$, and to JT ($\mu = 0.61$, $\sigma = 0.20$) with $p < 0.0001$ for session 2, comparing FT ($\mu = 0.22$, $\sigma = 0.13$) to CBCL ($\mu = 0.54$, $\sigma = 0.15$) with $p < 0.0001$ and to JT ($\mu = 0.48$, $\sigma = 0.19$) with $p = 0.0006$ for session 3, comparing FT ($\mu = 0.24$, $\sigma = 0.13$) to CBCL ($\mu = 0.53$, $\sigma = 0.20$) with $p < 0.0001$ and to JT ($\mu = 0.46$, $\sigma = 0.19$) with $p = 0.0018$ for session 4, and comparing FT ($\mu = 0.26$, $\sigma = 0.14$) to CBCL ($\mu = 0.55$, $\sigma = 0.19$) with $p < 0.0001$ and to JT ($\mu = 0.51$, $\sigma = 0.20$) with $p = 0.0010$ for session 5. No statistically significant difference is seen between JT and CBCL for any of the sessions (more details in Table 7). Further, when considering the two models separately, significant differences are seen between the first and the subsequent sessions for FT only.

As evident from the ANOVA, there was no statistically significant difference in classification accuracy for expert and non-expert users (based on their previous programming experience). Therefore, we did not perform a *post hoc* Tukey's HSD test for this variable.

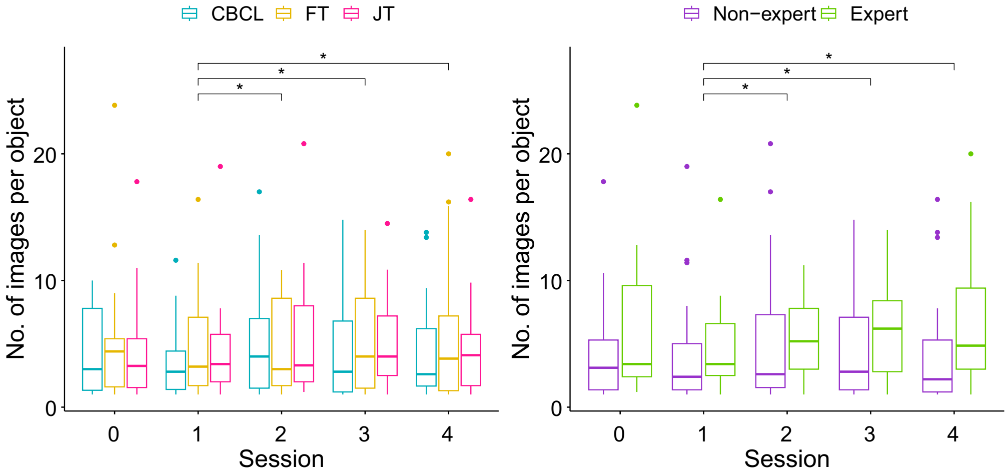
*5.2.2 Number of Images Per Object.* We performed the *post hoc* Tukey's HSD test for the significant ANOVAs for the number of images as the dependent variable. Figure 6 details the difference between the three CL models and expert and non-expert participants in terms of the number of images taught per object. In accordance with ANOVA, there is no statistically significant difference between the three conditions or between experts and non-experts for individual sessions and all sessions combined. Figure 7 correlates with the ANOVA. In terms of individual sessions, there is no statistically significant difference between the first and the second session. However, there is a statistically significant difference between session 2 ($\mu = 4.23$, $\sigma = 3.69$) and session 3 ($\mu = 5.14$,



(a) Number of images per object for experts and non-experts, and three conditions with CL methods (CBCL, FT, JT).

(b) Number of images per object by session for experts and non-experts in FT condition.

Fig. 6. Boxplots for *number of images per object*. Significance levels ($*p < .05$; $***p < 0.001$; $****p < 0.0001$) are indicated on bars between columns.



(a) Number of images per object over five sessions for, the three conditions with CL methods (CBCL, FT, JT).

(b) Number of images per object by session for experts and non-experts.

Fig. 7. Boxplots for *number of images per object* over five sessions. Significance levels ($*p < .05$) are indicated on bars between columns.

$\sigma = 4.27$) with $p = 0.0222$, session 2 and session 4 ($\mu = 5.00$, $\sigma = 3.98$) with $p = 0.0400$, and between session 2 and session 5 ($\mu = 4.90$, $\sigma = 4.55$) with $p = 0.0180$. Detailed results are shown in Table 8.

As we noticed the effect of prior programming experience in [4] in the FT condition and also observed a higher number of images for experts in the FT condition (Figure 7), we performed a Wilcoxon rank sum test [66] with false discovery rate correction [11] between experts and non-experts in the FT condition over five sessions. As displayed in Figure 6, there is a statistically significant difference between experts and non-experts for sessions 4 and 5 only, i.e., when comparing experts ($\mu = 9.05$, $\sigma = 4.41$) to non-experts ($\mu = 3.68$, $\sigma = 3.25$) with $p = 0.035$, $W = 10.5$

in session 4, and comparing experts ($\mu = 11.49$, $\sigma = 7.03$) to non-experts ($\mu = 3.14$, $\sigma = 2.49$) with $p = 0.035$, $W = 10.0$ in session 5.

*5.2.3 Number of Teaching Phases Per Session.* As the ANOVA for the number of teaching phases was not significant, we did not perform a *post hoc* Tukey's HSD test. Overall, 35 out of 60 participants had at least one session where they started more than one teaching phase with the robot. Overall, 100 out of 300 sessions had more than one teaching phase ranging from 2 to 11 teaching phases in a single session.

*5.2.4 Reteaching After Misclassification.* As the ANOVA for reteaching after misclassification was not significant, we did not perform a *post hoc* Tukey's HSD test. Overall, we noticed that 33 out of 60 participants retaught at least one object after it was misclassified by the robot during the testing phase. In total, there were 90 out of 300 sessions in which participants retaught misclassified objects with a maximum of 7 misclassified objects retaught in a session.

Note there were two different kinds of objects that participants could reteach after they were misclassified by the robot in a session: objects that were taught in the session, and objects that were taught in previous sessions. The above statistic only counts the reteaching of misclassified objects from the current session only, i.e., if an object taught in the previous sessions was misclassified and retaught in a session it is not covered in the above statistic. The ANOVA for this variable was significant for the session number and interaction between the three independent variables; therefore, we performed a *post hoc* Tukey's HSD test for the significant ANOVAs. Figure 9 shows the number of old objects retaught after misclassification for the three CL models. There is a statistically significant difference between session 1 ($\mu = 0.07$, $\sigma = 0.52$) and session 4 ($\mu = 0.23$, $\sigma = 0.72$) with $p = 0.0345$, and between session 1 and session 5 ($\mu = 0.17$, $\sigma = 0.56$) with $p = 0.0484$ (more details in Table 9). There is no statistically significant difference between any other sessions and between the three conditions.

As most participants did not reteach old objects, the value for this variable was mostly zero for the 300 sessions. Therefore, we also report the raw statistics for this dependent variable. Overall, there were only 18 sessions when participants retaught at least one object from the previous sessions, with a maximum number of 4 old objects taught in a session. In terms of the number of participants, only 11 out of 60 participants retaught objects from previous sessions in subsequent sessions.

Finally, we also conducted the ANOVA with combined retaught objects in a session and previous sessions, because many participants did not reteach old objects. Table 12 in Appendix shows the ANOVA results. These results confirm that the ANOVA for reteaching after misclassification for this combined data was not significant. Therefore, we did not perform a *post hoc* Tukey's HSD test.

### 5.3 Semantic Analysis

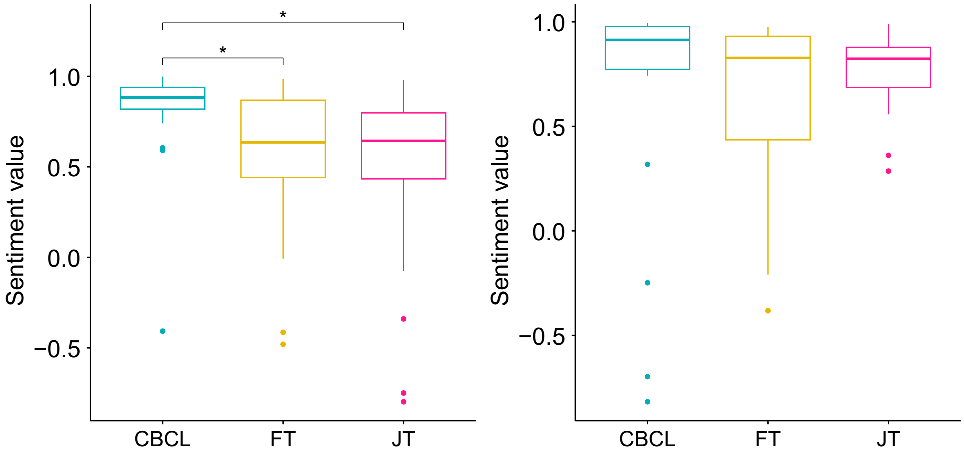
The overall sentiment scores for all questions are *Q1: robot_perception*: $\mu = 0.62, \sigma = 0.45$, *Q2: robot_house*: $\mu = 0.58, \sigma = 0.49$, *Q3: assist_friend*: $\mu = 0.68, \sigma = 0.42$, *Q4: concern_friend*: $\mu = 0.36, \sigma = 0.52$, *Q5: suggestion_improvement*: $\mu = 0.63, \sigma = 0.49$, *Q6: pre-defined_objects*: $\mu = 0.72, \sigma = 0.30$. In terms of the three conditions, we performed an ANOVA to determine if there was an effect of the choice of the CL model on sentiment scores for any of the six questions. Table 4 shows the $p$ values for all six open-ended questions. We performed *post hoc* Tukey's HSD tests for significant ANOVAs (Q1: robot_perception and Q3: assist_friend).

Figure 8 shows the difference in sentiment values between the three conditions for Q1: robot_perception. We notice a statistically significant difference between CBCL ($\mu = 0.81$, $\sigma = 0.29$) and FT ($\mu = 0.55$, $\sigma = 0.44$) with $p = 0.0417$, and between CBCL and JT ($\mu = 0.48$, $\sigma = 0.38$) with $p = 0.0435$. No statistically significant difference was observed between FT and JT. For Q3: assist_friend (Figure 8), *post hoc* Tukey's HSD did not show a statistically significant difference between the

Table 4. Results ($p$ Values) of the ANOVA for Open-Ended Questions with the Continual Learning Model as the Independent Variable

| Questions | $p$ |
|---|---|
| Q1: robot_perception | **0.0194** |
| Q2: robot_house | 0.1388 |
| Q3: assist_friend | **0.0499** |
| Q4: concern_robot | 0.0758 |
| Q5: suggestion_improvement | 0.0687 |
| Q6: pre-define_objects | 0.1081 |

Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are in bold.



(a) Compound sentiment values for three conditions in Q1. (b) Compound sentiment values for three conditions in Q3.

Fig. 8. Boxplots for compound sentiment values for the three conditions with different CL methods (CBCL, FT, and JT) in Q1: robot_perception and Q3: assist_friend. Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are indicated on bars between columns.

three conditions. This is in correlation with the borderline $p$ value in ANOVA for Q3: assist_friend. Overall summary statistics for all three conditions are shown in Table 10.

### 5.4 TA

We identified a total of 4 main themes and 24 categories among participants' responses. Table 5 shows the four themes, 24 categories, and the number of times they appeared in individual responses by the participants. Table 6 further shows the number of times each category appeared in participants' responses in the three conditions. The four themes with example responses for the corresponding categories are described next.

*5.4.1 Theme 1—Robot Interaction.* Thirty seven responses were about interaction with the robot. This theme describes participants' perceptions about interacting with the robot over repeated interactions. Almost all of the participants had a positive perception of the teaching process and interactions with the robot, many considering it to be easy, straightforward, and fun. The two categories in this theme along with example responses are described next.

Table 5. Categories for Responses to Open-Ended Interview Questions

| Theme | Code | Category | Count |
|---|---|---|---|
| Robot interaction | 1 | Operating the robot is straightforward. | 29 |
| | 2 | The robot is fun. | 8 |
| Robot usability | 3 | The robot can be used for older adults. | 31 |
| | 4 | The robot can be used for people with disabilities. | 27 |
| | 5 | The robot can be used for children. | 8 |
| | 6 | The robot can be used by anyone. | 3 |
| | 7 | I do not want the robot for myself. | 23 |
| | 8 | I do not want the robot for anyone. | 5 |
| | 9 | The robot can be used to find/fetch personal belongings. | 17 |
| | 10 | The robot can be used in the kitchen. | 7 |
| | 11 | The robot is useless. | 6 |
| Concerns | 12 | The robot is inaccurate/unreliable. | 92 |
| | 13 | The robot is forgetful. | 28 |
| | 14 | The robot is bulky. | 19 |
| | 15 | The robot's arm moves fast. | 6 |
| | 16 | The robot's arm moves slowly/rigidly. | 11 |
| | 17 | The robot is not safe. | 25 |
| Improvements | 18 | The robot needs better object recognition. | 20 |
| | 19 | The robot needs a better camera. | 17 |
| | 20 | The robot needs to keep information private. | 7 |
| | 21 | The robot should come with a pre-defined set of objects. | 29 |
| | 22 | The robot should not come with a pre-defined set of objects. | 11 |
| | 23 | Robust customization based on users' needs for pre-defined objects. | 15 |
| | 24 | The robot needs to be more human-like. | 11 |

Note that some categories were found in multiple responses of a single participant; therefore, their overall count is higher than $N = 60$.

1. *Operating the robot is straightforward.* Almost half of the participants mentioned that working with the robot was quite straightforward. Most of these participants specifically mentioned that the teaching process was "pretty easy" because they saw the demo only once and could teach the robot afterwards, e.g., "I thought the teaching process was very easy, it was very intuitive. You only showed me how to use the iPad once, and I didn't need to be reminded again during further sessions, and my sessions were relatively far apart, so it didn't seem to be difficult to remember at all.," or "I think the teaching process was very streamlined and straightforward. It doesn't get simpler than that.," or "It was very simple to do. Just showing the objects and then pressing the button to give it different views of the object." Another participant supported this view by commenting about the quick learning ability of the robot, "I would say it was pretty easy. Again, the robot is smart enough to grasp things really quickly. I would say he's a fast learner."

2. *The robot is fun.* Interactions with the robot during the teaching and testing phases were also considered to be fun by some participants, although most of the excitement was in the first session which went down later on. This was potentially due to the novelty effect. Participants' comments about their long-term interactions were positive about the beginning but neutral for the later sessions, "It was a lot of fun. I haven't interacted with that robot before. That was quite new. I think there was a lot of excitement in the first session that kind of went down.," or "Ok, so my experience was fun and a little bit boring in the middle but overall fun."

*5.4.2 Theme 2—Robot Usability.* One hundred twenty-seven responses were related to the usability of the robot. This theme describes participants' perceptions of the robot's usability in different domains. Most of the participants positively perceived the robot to be used for a

Table 6. Categories for Responses to Open-Ended Interview Questions for the Three Conditions with Different CL Methods (CBCL, FT, and JT)

| Theme | Category Code | CBCL | FT | JT |
|---|---|---|---|---|
| Robot interaction | 1 | 12 | 5 | 12 |
| | 2 | 5 | 2 | 1 |
| Robot usability | 3 | 10 | 9 | 12 |
| | 4 | 9 | 10 | 8 |
| | 5 | 4 | 0 | 4 |
| | 6 | 1 | 2 | 0 |
| | 7 | 8 | 6 | 9 |
| | 8 | 0 | 5 | 0 |
| | 9 | 8 | 6 | 3 |
| | 10 | 2 | 1 | 4 |
| | 11 | 3 | 1 | 2 |
| Concerns | 12 | 30 | 28 | 34 |
| | 13 | 1 | 22 | 5 |
| | 14 | 10 | 3 | 6 |
| | 15 | 4 | 1 | 1 |
| | 16 | 3 | 3 | 5 |
| | 17 | 9 | 8 | 8 |
| Improvements | 18 | 7 | 8 | 5 |
| | 19 | 10 | 5 | 2 |
| | 20 | 6 | 1 | 0 |
| | 21 | 7 | 12 | 10 |
| | 22 | 4 | 2 | 5 |
| | 23 | 7 | 3 | 6 |
| | 24 | 4 | 2 | 5 |

variety of assistive applications, whereas some found it to be useless either for themselves or everyone. The nine categories in this theme along with participants' example responses are described next.

3. *The robot can be used for older adults.* A significant number of participants perceived the robot to be useful for assisting older adults in their homes. Many of these participants commented about the potential use case of the robot for assisting their grandparents, "I think it can work, for example, for my grandmother, because she is sick and she cannot have some kind of physical activity and this robot can help her a lot." or more specifically to help them fetch objects, "The first I thought about was my grandma because she's very old. She's over 90. And it might help for reaching stuff, getting some stuff for her that she can't reach. And the robot knows where it is or can indicate it." or find objects, such as medicine, "I could see this being helpful with my grandmother, because she has dementia, and it could help tell different pill bottles or something apart so that she knows what to take or anything."

4. *The robot can be used for people with disabilities.* Similarly, a large number of participants perceived a more general use case of the robot for people with disabilities, and not just older adults who might have disabilities. Participants commented that the robot could be helpful for people with mobility disabilities to help fetch objects and medicine, "Yeah, people with mobility disabilities, I think the robot could be helpful for getting things for them." or "I definitely think it could be used

for people who need some sort of assistance if they have trouble finding things or getting things, maybe they have some sort of disability. I think it would really benefit them for them to be able to put an object and the robot to be able to find that object or they need their medication or things like that. I think it would be really beneficial for them." or "I can see it being used as an accessibility type of thing. So if somebody has trouble walking, the robot could get something for them."

5. *The robot can be used for children.* Some participants mentioned a unique use case of this robot for teaching children new objects, where the user could teach the robot personalized names for the objects and then the robot could be used to teach children, e.g., "I would say situations like there are some kids in the home like you have to teach them. Where is the mug? Where is the Sharpie? You can first ask the robot to learn stuff, and then the kid can play around with the robot and learn actual stuff. So for a kid, it would be like a toy." or "If you want to teach a small kid, like objects on the table, you could use this robot to teach objects because the kid would interact with the robot."

6. *The robot can be used by anyone.* A few participants perceived the robot to be useful for anyone, not just older adults or people with disabilities, e.g., "Yeah, I think anyone can use it." or more specifically to find objects in the home, "Sure, I can see it being used by almost everyone in my house, because it's great for me to find something but you can't really remember where you put it or you can't really see something that you've put somewhere. Also, if you're busy with something else, maybe you want to find out where something is and you're tied up with something else, maybe that can be a great help too."

7. *I do not want the robot for myself.* Although a large number of participants perceived the robot to be useful for older adults, people with disabilities, and children, many of them did not think that the robot could be useful for them, e.g., "I don't know. Like how much I would use it just because I don't know. I don't think that like I would find a use for it personally in my life." or "For me, I don't think I need a robot like this, but I think I can see this application and assistive care kind of environment." or "If it works, I think it's quite cool. Or maybe for me, just, I don't know if finding objects would be that helpful, but maybe for people in need, not able to move on their own. It might be quite helpful."

8. *I do not want the robot for anyone.* A small number of participants perceived the robot negatively and did not think that it could be used for anyone. Participants commented that they would not want their friends or relatives to use this robot, "Honestly, no, I would not suggest my friend or relative to have this robot.," others mentioned that it would be a waste of money, "I would try and convince someone I secretly hate to waste their money on the robot. Like, that is something I would kind of do as a person because it's like if they're rich enough to spend a bajillion dollars on such a robot. It's like, yes, you can waste your money on this robot."

9. *The robot can be used to find/fetch personal belongings.* Many participants also specifically mentioned that the robot can be used to fetch and find things in the home, although they did not specify it for any particular population group, e.g., "Yeah, it happens a lot that you need something and, for example, you feel lazy and you don't want to bring that stuff by yourself and this robot can be really helpful." or "I mean, I wouldn't mind if it goes and gets me things. Because you know when you like on your bed or something and you want to pick something up. So if it's going to go do it for me, yeah, that would be nice.," or to find things quickly, "Yeah, if you have a lot of objects and you're trying to find an object in your house, you can use this robot to find the object in a pretty less time."

10. *The robot can be used in the kitchen.* A few participants mentioned a specific use case for the robot in the kitchen. Participants commented that the robot can provide an extra hand when cooking things, "I could see some people using it, especially if it's something like, I don't know, maybe cooking, and they're like alone in the house, and they could just use an extra hand, even just like, hey, can you get this out of the cupboard, or where is this? So that could be useful for

them." or assist in a restaurant or a household "So immediate situation that comes to mind is used in a restaurant or use as an assistant robot to fetch objects in like kitchen household."

11. *The robot is useless.* A small number of participants perceived the robot to be useless, particularly in the real world where this system can be quite slow, e.g., "Actually my personal feeling is that it might not be so useful. I mean in real life it's really hard to use that in different situations and it was a little tiring, to be honest. You should wait a lot for the system to do your actual orders. So yeah, to be honest. I didn't like it very much. It's my personal feeling." or they commented that they could not think of any use case for the robot, "I don't have a proper use for it right now in mind, so I wouldn't recommend it to a relative or friend because I don't know where exactly it would be that useful to really take it."

*5.4.3 Theme 3—Concerns.* The most number of participants' responses (181) were related to concerns about the robot. This theme describes participants' concerns about the robot's deployment in the real world, mainly related to the reliability and safety of the robot. The six categories in this theme along with participants' example responses are described next.

12. *The robot is inaccurate/unreliable.* All of the participants had multiple concerns about the reliability of the robot, with many participants mentioning it multiple times in the open-ended questions. The concerns related to reliability were evenly distributed across the three conditions, and not only in the FT condition where the robot forgot previous objects. Participants commented that they got frustrated due to the unreliability of the robot, "I mean I kind of got frustrated at quite some points as a customer I know I can't rely on it right now. I don't know what errors it has but it does need some more things to be covered before introducing it to the market so." Others said that it was inconsistent and not 100% reliable, e.g., "It seems very inconsistent and its ability to even detect an object in the same spot is entirely random. Even if it's facing the exact same way or in the exact same position, it's very, very inconsistent.," or "The only concern I would have is that it's not 100% reliable all the time."

13. *The robot is forgetful.* A significant number of participants more specifically said that the robot forgot previous knowledge. Most of these participants were in the *FT* condition, which was expected as the FT model theoretically forgets all prior knowledge when learning new tasks. Participants commented that the robot was reliable for the objects taught in a session but forgot previous objects, "Well, I thought it was kind of inconsistent because it only recognizes the things that you teach immediately. I mean, like the five newest objects, and it forgets like the rest of them.," or "So, I thought that initially it was very easy to use and it was very reliable. But as more objects were added to its database, it started having a lot of problems identifying even objects that had no problem identifying before. And so initially you could have like five objects on the table and it'd have no problem finding them. But later on, sometimes even just having one object by itself was too difficult for the object. And then sometimes it works and sometimes it doesn't. Usually, it's able to learn the new object very well, but it kind of in a way forgets about the old items." Others commented that the robot was not able to remember previous objects, but they were not sure if it was because of the robot or their teaching style, "Well, the one obvious thing was that the robot had trouble rememberingthe objects that I had taught it, but I don't know if that was necessarily the robot or myself as a user."

14. *The robot is bulky.* Many participants had a concern regarding the robot's size, specifically that the robot was too bulky and big to be used in tight spaces in real-world environments. Participants commented that the robot might take up too much space and we might have to be careful not to bump into it, "Maybe it'll take up too much space and it needs a safe space around it. So yeah, maybe in the household it would be a little bit, everyone will have to be cautious around it so that they don't bump into it and you know, spoil the robot. Nothing will happen to us, but the robot

should not get damaged. So that could be a problem." or "It's a little bit huge. And I'm worried if it has a collision." Others commented that the robot will not be useful in smaller apartments, "Well, if you don't live in a bungalow, it's going to be hard to go downstairs if that needs to happen. It's kind of bulky so it might not be good in a Toronto apartment or a small apartment. Um, and its arm swinging around might not be very convenient in small spaces either."

15. *The robot's arm moves fast.* A few participants commented that the robot's arm moves really fast and it should be slowed down for safety, "Maybe move the arm a little bit more slowly because that kind of shows that it's being a little bit more careful because right now it might be a little bit too fast with the arm movement. You can't stop it either." Others commented that the pace of the robot might be too fast for some people, "Pretty much I have no concerns because it really worked well. It doesn't look like, you know, dangerous in all that. So maybe the pace of the robot might be too fast for people."

16. *The robot's arm moves slowly/rigidly.* In contrast, some participants had a concern about the slow pace of the robot and its arm. Participants commented to make the robot faster, but did appreciate that the slow movement makes it safe when interacting with people, "Just making sure that it's as reliable as humanly possible, and maybe a little bit faster, because it is a little bit slow in the movements, but also that's a good thing when it's interacting with people. You don't want it to be going super fast. I can see the pros and cons in speeding it up." Others mentioned that the robot's movements were too rigid, "Something that the robot developers should focus on would be to make the movements less rigid so it can move out of the way of obstacles."

17. *The robot is not safe.* Almost half of the participants had concerns regarding the safety of the robot, which were not only due to its arm speed or its size but also because of other factors. Participants commented that they do not really trust the robot when it moves especially when it comes close to some objects which they did not predict, "My concern is I don't trust it whenever it moves. I am not always sure on how it's going to move. When it points I have a very good idea, but there are a couple of times when it got really close to touching an object like the tall water bottle, and I'm excluding the time I did something dumb and put something where I shouldn't have, but it does sometimes get farther or closer to objects than I expect. And that arm looks like it could definitely break something. So I would be concerned if something like that was being used in my household. It would have to be very strictly restricted to not break anything." Others commented that the robot makes them scared and going too close to the robot might not be safe, e.g., "I was kind of scared when I saw it for the first time. So if there are children in a house they might get scared or we have an area, like a specific area for the robot to interact. So yeah, if people are too close to it, it might hurt people.," or "During this research, I have had to stand far away from the robot in the little squares on the floor, whenever it's moving. So, I guess if it was moving and bringing stuff around, it would have to be that I don't have to move out of its way, especially if I have a moving disability. So, I guess that would be a safety concern."

#### 5.4.4 Theme 4—Improvements.

The last theme is related to participants' comments to improve the robot. One hundred ten responses were related to the suggestions given by the participants for improving the robot so that it could be deployed in a household environment. The seven categories in this theme along with example responses by the participants are described next.

18. *The robot needs better object recognition.* Thirty percentage of the participants recommended improving the object recognition of the robot. Participants commented that the reliability of the object recognition should be improved "If its functions can be improved, mainly like being able to detect objects better and perform a little bit more reliably, I think it has applications in maybe larger than average household or for people with special needs.," others also mentioned to improve the robot to be able to remember previous objects, "I mean, I'm not a software or machine learning

expert, but I guess, there needs to be work on the remembering part, and then the recognition part needs to be worked on.," and robustness against changing angles of objects "I think one thing that can be improved is how it remembers objects from different angles because during the test, I tried to change the objects' angles or, shape or how it was put on the table during the teaching section, but after I changed it a little bit, the robot cannot recognize it."

19. *The robot needs a better camera.* A large number of participants also suggested improving the camera of the robot to improve object recognition. Participants commented to have multiple cameras for images from multiple angles, "I guess a better camera than this one so that it can recognize objects better. And maybe multiple cameras to see the object from multiple perspectives so that the user doesn't have to always put it in different orientations.," or take 3D images instead of 2D, "I would say first of all I am not sure what was happening because it couldn't detect several objects. So, I was thinking maybe right now it's just taking a 2D picture so I would like my robot to be taking a 3D picture so it is covering all the dimensions and the aspects of the objects. I think that will be helpful.," or improve the camera resolution "Maybe, increase the resolution of the camera. It seems like the camera resolution is not very good at the moment."

20. *The robot needs to keep information private.* Some participants had suggestions regarding data privacy, particularly because it collects users' personal data which should not be available to other people, e.g., "I think the main concern is the privacy that the data it collects must be completely secured and not used by any other person." or "Yeah, not much comes to mind except a wide berth and also maybe privacy concerns because it does take a look at the inside of my house and the objects in it and it has a camera I'm not sure when it's active or not so I don't know I'd be a bit wary."

21. *The robot should come with a pre-defined set of objects.* Almost half of the participants suggested that the robot should come with a pre-defined/trained set of objects so that the users do not have to teach it everything. Participants commented that having some pre-trained objects would ease working with the robot, e.g., "Yes, I would say having some pre-trained objects would be good because that will help the user, in general, to work with the robot easier.," or "I would definitely think that the developer should set up some stuff initially, just so that the robot has a certain level of understanding before just handing it off to the owner."

22. *The robot should not come with a pre-defined set of objects.* Contrarily, another set of participants suggested that the robot should not come with any pre-trained set of objects, and the users should teach it on their own. Participants commented that having pre-trained objects in the robot might not be useful for people, "I don't know if having it comes with pre-existing objects, I don't know what the use of those objects would be to those people, so I like this method better, where you can teach it whatever you want to fetch it.," others supported this by saying that everyone has a different set of objects so the users should teach the robot themselves, "It should not come with a predefined object, probably because everyone has a different set of objects. And then, someone calls the same objects different names, so it's better if one teaches their own style of objects.," and some commented that pre-defined objects could take up too much storage space, so the robot should only learn what they want it to learn, "The second option would be preferred like making him learn what we want him to learn rather than getting a predefined set of learned objects. Because of course, we don't want to consume a lot of storage."

23. *Robust customization based on users' needs for pre-defined objects.* Finally, another set of participants suggested that the choice of having pre-defined objects should be dependent on the users' needs, and even with pre-defined objects users should have the ability to customize and teach again. Participants commented that the pre-defined objects can depend on the application area, "I think it depends on how you want to use that robot. Yeah. For example, if it's asked for your household, then you might need to teach it everything you want to teach it. But if it's for a specific

experiment like this, then you can have predefined objects for it to learn.," others commented that there should be pre-defined objects but users should be able to train it with more objects if needed, e.g. "It should be a little bit of both. Maybe you can have the common items alone, like, already set. And then the user can train whatever other objects that they have," or "It should come with a predefined set of objects, but that should also be customizable by the user. So if the robot knows pop, but the user wants to differentiate between 7Up and Coke, they should be able to do that."

24. *The robot needs to be more human-like.* Some participants also suggested improving the interaction capabilities of the robot, to make it more engaging and user-friendly. Participants commented making it *human-like* in its appearance and the way it talked, e.g., "I think it must be a little, it must have a little human-like behavior in its intonation and in its appearance to make people more comfortable with it than people trust it better.," or "I don't know if it's possible, but having it be a little more engaging with the voice tone and a little more of a feel of like a human would be a better thing, especially in everyday household tasks and just being in the house, so it feels less like a robot." Others commented that making the robot human-like might be weird, but it should be user-friendly, "I don't know, maybe make it more human-like, but then that would be kind of weird, too, because it's a robot. It's not a human, but you know, kind of just like making it user-friendly."

Finally, we performed word cloud analysis on participants' responses to the open-ended questions. Results for the word cloud analysis are presented in Appendix B.

### 5.5 Teaching Experience Ratings

One of the open-ended questions asked participants to rate their experience teaching the robot on a scale of 1 (Very Boring) to 5 (Very Engaging). The overall rating for the teaching process was $\mu = 2.91, \sigma = 0.99$. We further performed a one-way ANOVA, but the results were not statistically significant with $p = 0.55$. Detailed summary statistics for the three conditions are shown in Table 11.

## 6 Discussion

Results from the qualitative and quantitative analyses of the data collected in our study allow us to validate the hypotheses in Section 5 and answer the research questions in Section 1.

### 6.1 Object Learning by Human Teachers (RQ1)

For object labeling, we noticed significant variations in the labeling strategies of different participants. None of the 25 objects used in the study had a single consistent label across 60 participants, even for simple objects, such as *Apple*, *Banana*, and so on. Further, some participants also label different objects with the same label, and some participants gave multiple labels to the same object. These strategies affected the performance of the CL robot for different participants as depicted by the high standard deviation in classification accuracy of the two CL models (Figure 5). As a consequence, we can accept *H1.1*: *Labeling strategies for objects vary among different users*, which answers *RQ1*: *How do different users label objects when teaching a CL robot over multiple sessions?* These results also indicate the need for developing personalized robots that adapt to users' labeling strategies and learn, and understand, their environment such that both the user and the robot can effectively communicate about the entities in the environment.

### 6.2 Robot's Performance (RQs 2, 3, and 4)

The classification accuracy of the CL robot was significantly affected by the choice of the CL model which was expected as the FT model forgets previous objects over the five sessions. However, classification accuracy was not affected by the previous robot programming experience of the participants. This result was surprising as it indicates that even expert users who have previous programming experience might not be familiar with CL over the long term. Therefore, the teaching effectiveness of both expert and non-expert users might be similar for a CL robot. This is promising for CL robots in the real world as most users will not have prior experience working with robots. Consequently, we have to reject *H4.1*: *CL robots taught by expert users perform better than the ones taught by non-expert users.*

Finally, both JT and CBCL achieve significantly higher accuracy when tested on data provided by the researchers in controlled experimental setups [8]. For example, CBCL and JT achieve ~90% accuracy after learning 22 different object classes over 11 sessions (2 classes per session) with only 5 training images per class. In contrast, in our experimental setup, both of these models achieve ~50% accuracy after 5 sessions. These results indicate that the results of CL models in controlled setups with researchers might not be the best indicator of real-world applicability, because participants' teaching styles in unconstrained experimental setups can significantly affect the performance of these models.

### 6.3 Participants' Teaching Styles (RQs 2, 3, and 4)

We quantified participants' teaching styles by calculating the number of images taught per object, the number of teaching phases started in each session, the number of times objects were retaught after being misclassified by the robot, and the number of times objects from previous sessions were taught by the participants. For the number of images per object, we did not find a statistically significant difference regarding the choice of the CL model and the previous programming experience of the participants. However, we did notice a statistically significant difference in the number of images shown per object over different sessions, with more images per object shown in later sessions. These results indicate that as the robot's performance might degrade over time, because of forgetting or other factors, users would compensate by spending more effort in teaching the robot. Further analysis also showed that expert users showed a significantly higher number of objects than non-experts in later sessions when interacting with the robot in the FT condition. This result further shows that based on their previous experiences expert users might try to compensate for the degraded performance of the robot in later sessions by spending more effort in teaching the robot. Note, that this might still not affect the robot's classification performance, as indicated by results for classification accuracy for expert and non-expert users. However, future studies with improved CL models might show differences in the performance of these models when learning from experts and non-experts.

For the number of teaching phases per session and reteaching of misclassified objects, there was no statistically significant effect of any of the three independent variables. We did, however, notice that more than 50% of the participants started multiple teaching phases per session with the robot. A similar result was also seen for reteaching misclassified objects. These results demonstrate the difference between constrained CL experimental setups and the real world where users might teach the robot multiple times in an interaction and also reteach the robot if it does not recognize an object correctly. In contrast, constrained CL setups train the robot to learn an object only once. Note that in the demo phase, participants were shown only a single teaching and testing phase with the robot, and they were not told that they could reteach misclassified objects to the robot. These results show that users might teach the robot differently than the experimenters or CL robot developers.

Finally, for reteaching old objects there was no statistically significant effect of the choice of the CL model or prior programming robot programming experience of the participants. This result is surprising because the FT model forgets the previous objects but CBCL and JT do not; therefore, we expected that participants in the FT condition would reteach more old objects than participants in the other two conditions. We did notice that participants taught a significantly higher number of old objects in later sessions. We believe the reason is that in the earlier sessions, there were no old objects to teach the robot and the robot performed reliably. However, by sessions 4 and 5, there are a large number of old objects taught in previous sessions, and the robot's performance also degrades because of forgetting. We further noticed that ~18% of the participants retaught old objects to the robot. These further confirm the difference between the constrained CL setups and real-world setups with robots interacting with real users, as we notice a natural replay of misclassified old information for the robot through user interactions. Note that in the study instructions, and during the demo phase, participants were not told that they cannot re-teach old objects. Similar to the results for the number of teaching phases and reteaching misclassified objects, these results show how real users might teach the CL robots. Finally, all of these results show that most users in the study were motivated to improve the performance of the robot, even though they were not given any specific incentive to do so. This is quite promising, as it indicates that users might be motivated to improve the performance of their personal robots over long-term interactions.

Based on the above results for the teaching style of the users, *H2.1* (*Classification performance of the robot affects the teaching style of the participants over multiple sessions*) can be accepted partially as we noticed that almost half of the participants retaught objects to the robot based on the robot's classification performance. We also noticed that experts taught a higher number of images per object to the robot in the FT condition in the later sessions when the robot's performance degraded. Both *H2.2* (*Users teach a robot that forgets previous objects differently than a robot that remembers previous objects*) and *H2.3* (*Users teach a robot that retrains on all previous objects differently than a robot that does not train on all previous objects*) have to be rejected as we did not notice the effect of the CL model on any of the variables representing the teaching style of the users. *H3.1* (*Teaching styles of users change over multiple sessions regardless of the CL model*) can be accepted partially as we did notice an effect of the sessions number on the number of images per object and reteaching old objects, but not on the number of teaching phases per session and reteaching misclassified objects. Finally, we did not see a statistically significant effect of the previous robot programming experience of the participants on any of the variables representing participants' teaching styles. Therefore, we have to reject *H4.2* (*There is a difference between the teaching styles of expert and non-expert users*). Overall, the above results and the hypotheses allow us to answer RQ2–4 regarding the teaching style of the users. Specifically, for *RQ2*: *How does the CL robot's performance affect the way users teach over multiple sessions?* we noticed a direct effect of model performance on participants' teaching styles as many participants re-taught objects to the robot after they were incorrectly classified. For *RQ3*: *Do users change the way they teach the CL robot over multiple sessions?* we noticed that participants did evolve their teaching styles over multiple sessions by teaching the robot more in later sessions. Finally, for *RQ4*: *Is there a difference in teaching style and robot performance for expert and non-expert users* we did not notice any significant effect on the teaching style or the robot's performance between expert and non-expert users.

## 6.4 Teaching Experience Ratings

Results for teaching experience ratings showed that participants were neither overly excited nor bored while interacting with the robot over multiple sessions. Further, the choice of the CL model did not have an effect on these ratings, indicating that the performance of the robot does not affect users' teaching experience with the robot. We hypothesize one of the reasons for low ratings was that the robot did not have any extra social cues in an overall functional setup. Utilizing social cues and changing the dialogue of the robot might improve users' overall experience with the robot over the long term.

### 6.5 Semantic Analysis (RQ5)

Results for semantic analysis of the open-ended questions showed an overall positive sentiment in all of the questions with $\mu > 0.5$, except for question 4 with $\mu = 0.36$. As Q4: concern_robot asked participants about their concerns about the robot, it was expected that sentiment values for many participants would be lower (neutral or negative sentiments), therefore the average sentiment value was lower for responses to this question. The standard deviation in the sentiment values for all the questions was also very high (~0.5) which shows that there were significant variations among the sentiment values for responses by different participants. Still, on average a positive sentiment for responses to all questions indicates that most participants had positive perceptions toward the robot. Further analysis indicated that the sentiment value in the responses to Q1: robot_perception (*What do you think about this robot or system?*) for participants in the CBCL condition was significantly higher than the sentiment values of responses for participants in the FT and JT conditions. It was expected that participants in the CBCL condition would perceive the robot more positively than participants in the FT condition because of the lower classification accuracy of the FT model, especially in later sessions. However, it was surprising to see that a similar effect was seen between the CBCL and JT conditions. We think, one possible explanation could be that the JT model requires a much higher amount of time to train on new objects in later sessions because it has to replay all the data from previous sessions, whereas CBCL can learn quickly as it does not require any replay. Therefore, this might have negatively affected participants' perceptions of the JT condition. These results indicate that users think more positively about robots that can learn quickly and perform reliably without forgetting old knowledge.

For the rest of the five questions, there was no effect of the choice of the CL model on the sentiment values in participants' responses. Both questions 2 and 3 asked participants about the perceived usability of the robot in household environments. The results indicate that participants had similar perceptions about the usability of the robot in household environments irrespective of the choice of the CL model. This was promising as it shows that even if the robot is forgetful or unreliable users perceive it to be similarly useful for household assistive applications. Questions 4: concern_robot, 5: suggestion_improvement, and 6: pre-define_objects asked participants about their concerns and suggestions for the future development of the CL robot. Again, results for these questions show that participants had similar sentiments in terms of concerns and suggestions toward all CL models. This indicates that CL robots, even if they do not forget past knowledge, are quite far from being deployed in real environments.

### 6.6 TA (RQ5)

TA of the open-ended questions resulted in four major themes indicating participants' perceptions of interaction with the robot, the usability of the robot for different real-world scenarios, concerns about the robot assisting with daily tasks, and suggestions for improving the robot for deployment in household environments. These themes were divided into 24 categories that provide further insight into the strengths and weaknesses of the CL robot. Almost half of the participants mentioned that it was straightforward to interact with, teach, and test the robot. Most of these participants were from the CBCL and JT conditions, which could be because participants in the FT condition had to reteach the robot multiple times after misclassifications, and therefore did not perceive the teaching process to be easy. In terms of usability, although many participants did not want the robot for themselves, a large number of participants mentioned that the robot could be useful for older adults, people with disabilities, and for fetching everyday personal belongings. A few participants also highlighted the usefulness of the robot for teaching children. A small number of participants did mention that the robot is not useful for anyone, although most of these participants were in the

FT condition. Overall, these results were quite promising as participants can envision CL robots as helpful assistants in household environments.

Participants also identified many concerns related to the CL robot for assisting with daily tasks. The primary concern raised by all participants (multiple times by some participants) was that the robot was unreliable and inaccurate. This correlates with the classification accuracy results for the three CL models where the average accuracy was never higher than ~50–60%. The other major concern was that the robot forgot objects from previous sessions, although almost all of the participants who raised this concern were from the FT condition. This was expected as the FT model forgets the previous knowledge when learning new information. The rest of the concerns were related to the safety of the robot, especially when it moves its arm. These concerns indicate that CL robots are currently not reliable and safe enough to be deployed in our daily environments. Additionally, we noticed that there were conflicting comments by the participants regarding the robot's arm's speed, with some saying it was too fast and others saying it was too slow. Future research in CL should also focus on the relation between the perception of robots and safety [27] and personalization to users' preferences of perceived safety [65] in addition to learning task knowledge.

Finally, participants highlighted various ways to improve the robot for deployment in household environments. In correlation with the major concerns, the main suggestions for improvement were related to the object recognition and camera of the robot. Future development of CL robots should focus on improving the performance of the robot through robust CL methods. In response to the last open-ended question, half of the participants mentioned that it is better if the robot comes with pre-defined objects without any teaching required by the users. The rest of the participants were divided into two groups. One of them mentioned that the robot should not come with pre-defined objects and that users should be allowed to teach what they want the robot to learn. The second group mentioned a hybrid approach where pre-training or no-pretraining can be chosen based on the application area, e.g., for household applications users should be allowed to teach the robot their own objects. These results indicate that the development of CL robots should be based on their application area, and for household assistive applications, it might be useful to develop robots that can personalize to their users' environments through direct learning from the users. Overall, these results allow us to answer *RQ5*: *How do human users perceive the CL robot for everyday applications?*. Our analysis shows that users did perceive the robot to be useful in the future for assisting older adults and people with disabilities in their homes, albeit after several improvements related to the robot's object recognition, its camera, its interaction with the participants, and data privacy.

## 7 Conclusions

In this article, we considered a human-centered approach to CL to understand how users interact with, teach, and perceive CL robots over the long term. We designed a long-term between-participant HRI study in which 60 participants interacted with a CL robot using three different CL models (three conditions) over 300 sessions. We analyzed the data collected from this study to understand the different teaching styles of participants, and how these styles are influenced by the performance of the robot over multiple sessions. Our results indicate that different users teach household objects to the CL robot in a variety of ways, and, unlike constrained CL test setups, the classification performance of the robot can also influence the teaching style of the users. The results also show that the previous programming experiences of the users did not affect their teaching style or the performance of the robot, which is promising because it demonstrates that CL robots can easily learn from non-expert users. Our analysis of the open-ended questionnaires with the participants provides further insight into the strengths and weaknesses of CL robots and their potential to be deployed as assistive robots for older adults and people with disabilities in household environments. Overall, based on the results of this study, we recommend future CL models focus on adapting to the teaching style of their users, and that CL models should be tested in more realistic test setups.

## 8 Limitations and Future Work

We conducted our study in an unconstrained setup, where participants could teach and test the robot flexibly. However, the study was conducted in a robotics lab and not in a realistic household environment. In future work, we plan to conduct a similar study in a smart home with the same robot to understand the influence of the household environment on the interactions and teaching styles of the users.

We conducted the user study with a mix of expert and non-expert users, however, due to practical constraints of recruitment, they were all university students between the ages of 18 and 37 years. In future work, we plan to conduct this study with participants who might be less familiar with robots to understand the effectiveness of CL robots for assistive applications. Finally, the study was conducted with one particular robot and with three CL models. Expanding this work to other robots and CL models can help us understand the larger design space of CL robots and users' teaching patterns when interacting with these robots.

The GUI used in the system currently requires participants to type the name of the object during the training and testing phases. This can be cumbersome, and people can also forget the names they gave to the objects over time. In the future, one research direction can be to improve the GUI and the interaction with the robot, such as by saving the names of the objects that the participants provide over time and then providing an option to choose an object name from a drop-down menu rather than typing it every time. Additionally, adding an NLP-based interaction supporting the GUI or standalone could also improve how users interact with and teach the robot. This could be effective for older adults and people with disabilities who might be unfamiliar with technology and face difficulties interacting through a GUI. Future studies with an NLP-based interaction system, particularly with older adults, could shed better light on the capabilities of this system for real-world assistive applications.

Despite these limitations, our user study took the first step toward a human-centered approach to CL by integrating ML-based CL models with HRI. We hope that our results can help ML and HRI researchers design CL models that can adapt to their users' teaching styles and test these models in realistic experimental setups where embodied agents interact with human users.

## Appendices

## A Detailed Results of the Study

Tables 7–9 show detailed summary stats and $p$ values for the classification accuracy of the three CL models, number of images per object, and number of old objects retaught by the participants, respectively. Table 10 shows the complete summary stats for the sentiment values in the responses to the open-ended questions for the three conditions.

## B Word Cloud Analysis

We performed word cloud analysis for responses to each of the six questions using the Python wordcloud library. Note that we filtered out common words in responses that were not helpful, such as "the," "like," and so on. Figure 10 shows word clouds for the six questions. The word clouds' most common words match the most frequent categories in the TA. For example, the most common

Table 7. Detailed Results for Classification Accuracy in the Three Conditions with Different Continual Learning Methods (CBCL, FT, and JT)

| Session | CBCL | | FT | | JT | | FT-CBCL | CBCL-JT | FT-JT |
|---|---|---|---|---|---|---|---|---|---|
| Value | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $p$ | $p$ | $p$ |
| 1 | 0.53 | 0.18 | 0.51 | 0.18 | 0.59 | 0.18 | 0.9777 | 0.4770 | 0.4052 |
| 2 | 0.54 | 0.16 | 0.29 | 0.13 | 0.61 | 0.2 | **0.0004** | 0.5291 | **<0.0001** |
| 3 | 0.54 | 0.15 | 0.22 | 0.13 | 0.48 | 0.2 | **<0.0001** | 0.4154 | **0.0006** |
| 4 | 0.53 | 0.2 | 0.24 | 0.13 | 0.46 | 0.19 | **<0.0001** | 0.6974 | **0.0018** |
| 5 | 0.55 | 0.19 | 0.26 | 0.14 | 0.51 | 0.2 | **<0.0001** | 0.7763 | **0.0010** |

Significance levels ($*p < .05; **p < 0.01; ***p < 0.001; ****p < 0.0001$) are in bold.

Table 8. Detailed Results for *Number of Images Per Object* in Five Sessions

| Session | | | 1-x | 2-x | 3-x | 4-x |
|---|---|---|---|---|---|---|
| Value | $\mu$ | $\sigma$ | $p$ | $p$ | $p$ | $p$ |
| 1 | 4.73 | 4.37 | - | - | - | - |
| 2 | 4.23 | 3.69 | 0.3489 | - | - | - |
| 3 | 5.14 | 4.27 | 0.9123 | **0.0222** | - | - |
| 4 | 5.00 | 3.98 | 0.8583 | **0.0400** | 0.9998 | - |
| 5 | 4.91 | 4.55 | 0.7591 | **0.0180** | 0.9968 | 0.9993 |

The column labeled 1-x shows $p$ values between session 1 and the rest of the four sessions. The next three column labels follow the same procedure. Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are in bold.

Table 9. Detailed Results for *Number of Old Objects* Retaught in Five Sessions

| Session | | | 1-x | 2-x | 3-x | 4-x |
|---|---|---|---|---|---|---|
| Value | $\mu$ | $\sigma$ | $p$ | $p$ | $p$ | $p$ |
| 1 | 0 | 0 | - | - | - | - |
| 2 | 0.07 | 0.52 | 0.9179 | - | - | - |
| 3 | 0.1 | 0.49 | 0.4191 | 0.9397 | - | - |
| 4 | 0.23 | 0.72 | **0.0345** | 0.2491 | 0.4719 | - |
| 5 | 0.17 | 0.56 | **0.0484** | 0.2625 | 0.7138 | 0.8589 |

The column labeled 1-x shows $p$ values between session 1 and the rest of the four sessions. The next three column labels follow the same procedure. Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are in bold.

words in Q3's word cloud are *useful*, *recommend*, *relatives*, *older*, *helpful*, *fetch*, and *home*. These match with common categories in the theme about the usefulness of the robot, i.e., the robot is useful for older adults, and the robot can be used to fetch personal belongings.

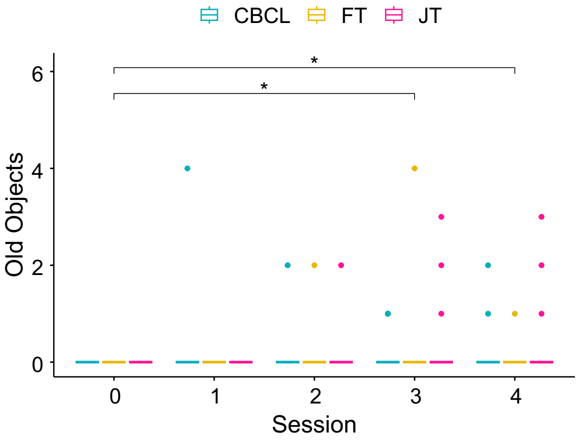
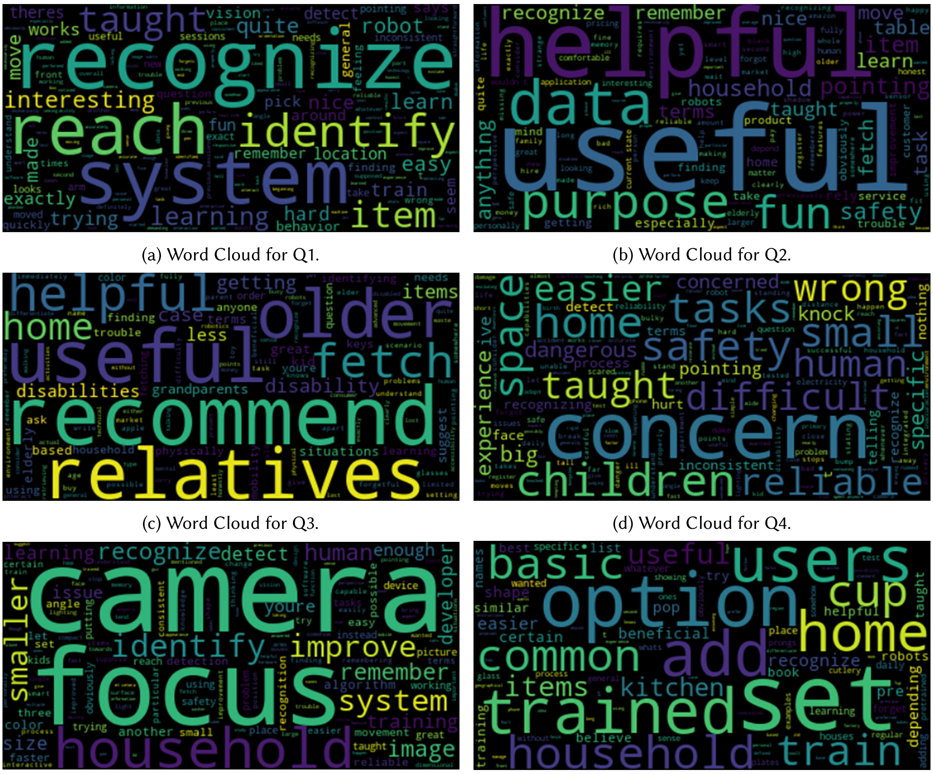


Fig. 9. Boxplot for *number of old objects retaught after misclassification* for the three conditions with different CL methods (CBCL, FT, and JT). Significance levels ($**p < 0.01$) are indicated on bars between columns.



(a) Word Cloud for Q1.

(b) Word Cloud for Q2.

(c) Word Cloud for Q3.

(d) Word Cloud for Q4.

(e) Word Cloud for Q5.

(f) Word Cloud for Q6.

Fig. 10. Word clouds for participants' responses to open-ended questions.

Table 10. Detailed Summary Statistics for Compound Sentiment Values for the Three Conditions with Different Continual Learning Methods (CBCL, FT, and JT) in Six Open-Ended Questions

| | CBCL | | FT | | JT | |
|---|---|---|---|---|---|---|
| Questions | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Q1: robot_perception | 0.81 | 0.29 | 0.55 | 0.44 | 0.48 | 0.55 |
| Q2: robot_house | 0.69 | 0.41 | 0.66 | 0.45 | 0.38 | 0.58 |
| Q3: assist_friend | 0.67 | 0.55 | 0.61 | 0.43 | 0.76 | 0.19 |
| Q4: concern_robot | 0.51 | 0.48 | 0.41 | 0.49 | 0.15 | 0.53 |
| Q5: suggestion_improvement | 0.81 | 0.36 | 0.54 | 0.53 | 0.53 | 0.57 |
| Q6: pre-defined_objects | 0.75 | 0.30 | 0.82 | 0.14 | 0.61 | 0.38 |

Table 11. Detailed Summary Statistics for Teaching Experience Ratings for the Three Conditions with Different Continual Learning Methods (CBCL, FT, and JT)

| CL Model | $\mu$ | $\sigma$ |
|---|---|---|
| CBCL | 3.05 | 1.10 |
| FT | 2.71 | 1.00 |
| JT | 2.96 | 0.86 |

Table 12. Results ($p$ Values and Effect Sizes) of the Three-Way ANOVA Using Session Number, Continual Learning Model, and Previous Programming Experience as Independent Variables

| *p Values* | | | | |
|---|---|---|---|---|
| | **Accuracy** | **Number of Images** | **Teaching Phases** | **Reteaching** |
| Programming experience | 0.5772 | 0.0845 | 0.4399 | 0.9718 |
| Session number | **<0.0001** | **0.0257** | 0.3903 | 0.2691 |
| CL model | **<0.0001** | 0.3748 | 0.8394 | 0.2404 |
| Programming experience: Session number | 0.5607 | 0.0534 | 0.5083 | 0.4787 |
| Programming experience: CL model | 0.856 | 0.0868 | 0.6896 | 0.5679 |
| CL model: Session number | **0.0011** | 0.2624 | 0.6151 | 0.4890 |
| Programming experience: CL model: Session number | 0.9111 | 0.0993 | 0.3127 | 0.1344 |

| *Effect Sizes* | | | | |
|---|---|---|---|---|
| | **Accuracy** | **Number of Images** | **Teaching Phases** | **Reteaching** |
| Programming experience | 0.0027 | 0.0424 | 0.0091 | <0.0001 |
| Session number | 0.0714 | 0.008 | 0.0034 | 0.0129 |
| CL model | 0.2467 | 0.0267 | 0.0053 | 0.0243 |
| Programming experience: Session number | 0.007 | 0.0065 | 0.0026 | 0.0085 |
| Programming experience: CL model | 0.0027 | 0.0702 | 0.0112 | 0.0096 |
| CL model: Session number | 0.0674 | 0.0063 | 0.005 | 0.0181 |
| Programming experience: CL model: Session number | 0.0074 | 0.009 | 0.0079 | 0.0313 |

Columns for the accuracy of the models, number of images per object, number of teaching phases, and reteaching misclassified objects (including retaught old objects) show $p$ values (top table) and effect sizes (bottom table) for the dependent variables. We use generalized eta [39] to calculate effect sizes. Significance levels ($*p < .05$; $**p < 0.01$; $***p < 0.001$; $****p < 0.0001$) are in bold.

## References

[1] 2005. Qualtrics. Retrieved from https://www.qualtrics.com

[2] Ali Ayub, Zachary De Francesco, Patrick Holthaus, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2024. Continual Learning through Human-Robot Interaction - Human Perceptions of a Continual Learning Robot in Repeated Interactions. *International Journal of Social Robotics*, under review. arXiv:2305.16332. Retrieved from https://arxiv.org/abs/2305.16332

[3] Ali Ayub and Carter Fendley. 2022. Few-Shot Continual Active Learning by a Robot. In *Advances in Neural Information Processing Systems.* Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). Retrieved from https://openreview.net/forum?id=35I4narr5A

[4] Ali Ayub, Jainish Mehta, Zachary De Francesco, Patrick Holthaus, Kerstin Dautenhahn, and Chrystopher L. Nehaniv. 2023. How Do Human Users Teach a Continual Learning Robot in Repeated Interactions?. In *Proceedings of IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* 1975–1982.

[5] Ali Ayub, Chrystopher L Nehaniv, and Kerstin Dautenhahn. 2022. Don't Forget to Buy Milk: Contextually Aware Grocery Reminder Household Robot. In *Proceedings of IEEE International Conference on Development and Learning (ICDL'22).* IEEE, 299–306.

[6] Ali Ayub and Alan Wagner. 2021. EEC: Learning to Encode and Regenerate Images for Continual Learning. In *Proceedings of the International Conference on Learning Representations*.

[7] Ali Ayub and Alan R. Wagner. 2020. Cognitively-Inspired Model for Incremental Learning Using a Few Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 222–223.

[8] Ali Ayub and Alan R. Wagner. 2020. Tell Me What This is: Few-Shot Incremental Object Learning by a Robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 8344–8350.

[9] Ali Ayub and Alan R. Wagner. 2020. What Am I Allowed to Do Here?: Online Learning of Context-Specific Norms by Pepper. In *Social Robotics.* Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge (Eds.), Springer International Publishing, Cham, 220–231.

[10] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. 2018. Learning from Physical Human Corrections, One Feature at a Time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 141–149.

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[12] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. 2022. Doodle It Yourself: Class Incremental Learning by Drawing a Few Sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2293–2302.

[13] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

[14] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D. Dragan. 2021. Feature Expansive Reward Learning: Rethinking Human Input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21).* ACM, New York, NY, 216–224. DOI: https://doi.org/10.1145/3434073.3444667

[15] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: https://doi.org/10.1191/1478088706qp063oa

[16] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-End Incremental Learning. In *Proceedings of the European Conference on Computer Vision (ECCV).* 233–248.

[17] Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to Action: Towards Interactive Task Learning with Physical Agents. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18.* International Joint Conferences on Artificial Intelligence Organization, 2–9. DOI: https://doi.org/10.24963/ijcai.2018/1

[18] Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2018. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV).* 532–547.

[19] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-Shot Classification. In *Proceedings of the International Conference on Learning Representations*.

[20] Vivienne Bihe Chi and Bertram F. Malle. 2023. People Dynamically Update Trust When Interactively Teaching Robots. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23).* ACM, New York, NY, 554–564. DOI: https://doi.org/10.1145/3568162.3576962

[21] Kerstin Dautenhahn. 2007. Socially Intelligent Robots: Dimensions of Human–Robot Interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362, 1480 (2007), 679–704.

[22] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand. 2019. Online Object and Task Learning via Human Robot Interaction. In *Proceedings of the International Conference on Robotics and Automation (ICRA'19)*. 2132–2138.

[23] Robert M. French. 2019. Dynamically Constraining Connectionist Networks to Produce Distributed, Orthogonal Representations to Reduce Catastrophic Interference. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society.* 335–340.

[24] Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. 2020. REMIND Your Neural Network to Prevent Catastrophic Forgetting. In *Computer Vision – ECCV 2020.* Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.), Springer International Publishing, Cham, 466–483.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770–778.

[26] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. 2022. Constrained Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9057–9067.

[27] Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. 2019. How a Robot's Social Credibility Affects Safety Performance. In *International Conference on Social Robotics (ICSR'19).* Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro-González, and Hongsheng He (Eds.), Lecture Notes in Computer Science, Vol. 11876. Springer Cham, 740–749. DOI: https://doi.org/10.1007/978-3-030-35888-4_69

[28] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. 2017. Learning Non-Maximum Suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[29] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 831–839.

[30] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F. Glas, Iolanda Leite, and Kheng Lee Koay. 2019. Personalization in Long-Term Human-Robot Interaction. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. 685–686. DOI: https://doi.org/10.1109/HRI.2019.8673076

[31] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Sinan Kalkan, German I. Parisi, and Hatice Gunes. 2021. Lifelong Learning and Personalization in Long-Term Human-Robot Interaction (Leap-HRI). In *Proceedings of the Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 724–727.

[32] Minsoo Kang, Jaeyoo Park, and Bohyung Han. 2022. Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16071–16080.

[33] Tasneem Kaochar, Raquel Torres Peralta, Clayton T. Morrison, Ian R. Fasel, Thomas J. Walsh, and Paul R. Cohen. 2011. Towards Understanding How Humans Teach Robots. In *Proceedings of the User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011*, Proceedings 19. Springer, 347–352.

[34] Ronald Kemker and Christopher Kanan. 2018. FearNet: Brain-Inspired Model for Incremental Learning. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=SJ1Xmf-Rb

[35] Cory D. Kidd and Cynthia Breazeal. 2008. Robots at Home: Understanding Long-Term Human-Robot Interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 3230–3235.

[36] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming Catastrophic Forgetting in Neural Networks. In *Proceedings of the National Academy of Sciences of the United States of America* 114, 13 (2017), 3521–3526.

[37] Kheng Lee Koay, Matt Webster, Clare Dixon, Paul Gainer, Dag Syrdal, Michael Fisher, and Kerstin Dautenhahn. 2021. Use and Usability of Software Verification Methods to Detect Behaviour Interference When Teaching an Assistive Home Companion Robot: A Proof-of-Concept Study. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 402–422.

[38] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially Situated Artificial Intelligence Enables Learning from Human Interaction. In *Proceedings of the National Academy of Sciences* 119, 39 (2022), e2115730119.

[39] Daniel Lakens. 2013. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology* 4, 1664–1078

[40] Y. LeChun. 1998. The Mnist Database of Handwritten Digits. Retrieved from http://yann.lecun.com/exdb/mnist/

[41] Z. Li and D. Hoiem. 2018. Learning Without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (Dec. 2018), 2935–2947.

[42] Vincenzo Lomonaco and Davide Maltoni. 2017. CORe50: a New Dataset and Benchmark for Continuous Object Recognition. In *Proceedings of the First Annual Conference on Robot Learning*, Vol. 78. 17–26.

[43] Maja J. Matarić. 2017. Socially Assistive Robotics: Human Augmentation Versus Automation. *Science Robotics* 2, 4 (2017). 5410.

[44] James L. Mcclelland, Bruce L. Mcnaughton, and Randall C. O'Reilly. 1995. Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review* 102, 3 (1995), 419–457. DOI: https://doi.org/10.1037/0033-295x.102.3.419

[45] Martin Mundt, Steven Lang, Quentin Delfosse, and Kristian Kersting. 2022. CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rHMaBYbkkRJ

[46] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. 2019. Learning to Remember: A Synaptic Plasticity Driven Framework for Continual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11321–11329.

[47] Laura Petrecca. 2018. How Robot Caregivers Will Help an Aging U.S. Population. Retrieved from https://www.aarp.org/caregiving/home-care/info-2018/new-wave-of-caregiving-technology.html

[48] Preeti Ramaraj, Charles L. Ortiz, and Shiwali Mohan. 2021. Unpacking Human Teachers' Intentions for Natural Interactive Task Learning. In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1173–1180.

[49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Ulrich Reiser, Theo Jacobs, Georg Arbeiter, Christopher Parlitz, and Kerstin Dautenhahn. 2013. Care-O-bot® 3 – Vision of a Robot Butler. *Your Virtual Butler* 7407, 97–116.

[52] Astrid Marieke Rosenthal-von der Pütten, Nikolai Bock, and Katharina Brockmann. 2017. Not Your Cup of Tea? How Interacting with a Robot can Increase Perceived Self-Efficacy in HRI and Evaluation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 483–492.

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (Dec. 2015), 211–252.

[54] R. M. Sakia. 1992. The Box-Cox Transformation Technique: A Review. *The Statistician* 41, 2 (1992), 169. DOI: https://doi.org/10.2307/2348250

[55] Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. 2016. "Teach Me–Show Me"—End-User Personalization of a Smart Home and Companion Robot. *IEEE Transactions on Human-Machine Systems* 46, 1 (2016), 27–40.

[56] Emmanuel Senft, Paul Baxter, James Kennedy, Séverin Lemaignan, and Tony Belpaeme. 2017. Supervised Autonomy for Online Learning in Human-Robot Interaction. *Pattern Recognition Letters* 99, 77–86.

[57] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual Learning with Deep Generative Replay. *Advances in Neural Information Processing Systems* 30, 2990–2999.

[58] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2021. Always Be Dreaming: A New Approach for Data-Free Class-Incremental Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9374–9384.

[59] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. 2020. Topology-Preserving Class-Incremental Learning. In *Proceedings of the Computer Vision – ECCV 2020: 16th European Conference*, Proceedings, Part XIX. Springer-Verlag, 254–270.

[60] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. 2020. Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12183–12192.

[61] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable Robots: Understanding Human Teaching Behavior to Build More Effective Robot Learners. *Artificial Intelligence* 172, 6–7 (2008), 716–737.

[62] Andrea L. Thomaz and Maya Cakmak. 2009. Learning About Objects with Human Teachers. In *Proceeding of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI'09)*. 15–22.

[63] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. 2023. A Survey on Few-Shot Class-Incremental Learning. arXiv:2304.08130. Retrieved from https://arxiv.org/abs/2304.08130

[64] Sepehr Valipour, Camilo Perez Quintero, and Martin Jägersand. 2017. Incremental Learning for Robot Perception Through HRI. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'17)*. 2772–2777.

[65] Sanne van Waveren, Rasmus Rudling, Iolanda Leite, Patric Jensfelt, and Christian Pek. 2023. Increasing Perceived Safety in Motion Planning for Human-Drone Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (Stockholm, Sweden) (HRI '23)*. ACM, New York, NY, 446–455. DOI: https://doi.org/10.1145/3568162.3576966

[66] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. Retrieved from http://www.jstor.org/stable/3001968

[67] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. 2016. Fetch and Freight: Standard Platforms for Service Robot Applications. In *Proceedings of the IJCAI, Workshop on Autonomous Mobile Service Robots*. 1–6.

[68] Chenshen Wu, Luis Herranz, Xialei Liu, yaxing wang, Joost van de Weijer, and Bogdan Raducanu. 2018. Memory Replay GANs: Learning to Generate New Categories without Forgetting. *Advances in Neural Information Processing Systems* 31, 5962–5972.

[69] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large Scale Incremental Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 374–382.

[70] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. 2021. Few-Shot Incremental Learning With Continually Evolved Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12455–12464.

Received 1 November 2023; revised 9 February 2024; accepted 28 March 2024