QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge

Sherry Ruan
ssruan@stanford.edu
Stanford University

Liwei Jiang
ljiang@colby.edu
Colby College

Justin Xu
justinx@stanford.edu
Stanford University

Bryce Joe-Kun Tham
bjtham@stanford.edu
Stanford University

Zhengneng Qiu
qiuzhengneng@gmail.com
Stanford University

Yeshuang Zhu
zhu-ys13@mails.tsinghua.edu.cn
Tsinghua University

Elizabeth L. Murnane
emurnane@stanford.edu
Stanford University

Emma Brunskill
ebrun@cs.stanford.edu
Stanford University

James A. Landay
landay@stanford.edu
Stanford University

ABSTRACT

Advances in conversational AI have the potential to enable more engaging and effective ways to teach factual knowledge. To investigate this hypothesis, we created QuizBot, a dialogue-based agent that helps students learn factual knowledge in science, safety, and English vocabulary. We evaluated QuizBot with 76 students through two within-subject studies against a flashcard app, the traditional medium for learning factual knowledge. Though both systems used the same algorithm for sequencing materials, QuizBot led to students recognizing (and recalling) over 20% more correct answers than when students used the flashcard app. Using a conversational agent is more time consuming to practice with; but in a second study, of their own volition, students spent 2.6x more time learning with QuizBot than with flashcards and reported preferring it strongly for casual learning. Our results in this second study showed QuizBot yielded improved learning gains over flashcards on recall. These results suggest that educational chatbot systems may have beneficial use, particularly for learning outside of traditional settings.

CCS CONCEPTS

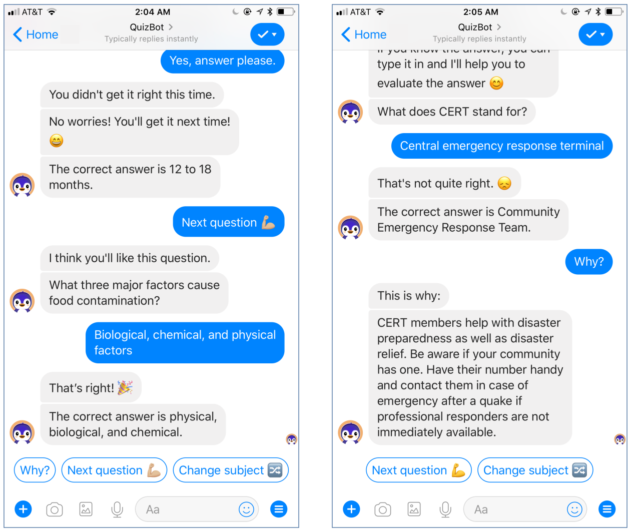
• **Human-centered computing** → **Natural language interfaces**; • **Applied computing** → **Education**; • **Computing methodologies** → *Artificial intelligence.*

ACM ISBN 978-1-4503-5970-2/19/05.
https://doi.org/10.1145/3290605.3300587

KEYWORDS

educational applications, pedagogical agents, chatbots, conversational user interfaces, intelligent systems

ACM Reference Format:
Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300587



Yes, answer please.

You didn't get it right this time.

No worries! You'll get it next time! 😄

The correct answer is 12 to 18 months.

Next question 💪

I think you'll like this question.

What three major factors cause food contamination?

Biological, chemical, and physical factors

That's right! 🎉

The correct answer is physical, biological, and chemical.

type it in and I'll help you to evaluate the answer 😊

What does CERT stand for?

Central emergency response terminal

That's not quite right. 😞

The correct answer is Community Emergency Response Team.

Why?

This is why:

CERT members help with disaster preparedness as well as disaster relief. Be aware if your community has one. Have their number handy and contact them in case of emergency after a quake if professional responders are not immediately available.

Figure 1: Screenshots of QuizBot: (left) a user just answered a question correctly by typing an answer, (right) feedback was given to a user's incorrect answer.

1 INTRODUCTION

A great deal of learning involves factual knowledge (e.g., numerous topics in medicine, language, and law). Further, such information is often learned outside of a formal classroom setting. Developing more effective automated methods for accelerating or improving factual learning therefore has the potential to benefit a multitude of students on a broad scale.

Traditional electronic tools for practicing factual knowledge tend to be flashcard based [15, 29, 30]. Flashcard apps are simple and can be easily designed to provide personalized adaptive practice based on well-studied models of human memory [14, 50, 55]. However, to optimize for speed, flashcards typically involve passive learning (i.e., the user is asked to visualize the answer and then check for correctness). This may not fully take advantage of the *testing effect* (retrieval through testing with proper feedback) [48]. As shown in many previous studies, retrieval practices like testing lead to higher retention than purely studying via even multiple passive means of self-evaluation [8, 34, 47]. Feedback received from test results further improves retention [3, 37].

Moreover, flashcards are not typically designed to be engaging, making their effectiveness heavily dependent on people's desire to learn. Research confirms engagement can mediate learning effectiveness [7, 27], especially for technology-based learning [26]. A more engaging way to learn factual knowledge could therefore lead to better learning outcomes.

One possible path towards boosted engagement is using Natural Language Processing (NLP) powered chatbots, which are becoming increasingly sophisticated [21, 52]. For example, such systems enable students to speak or type out their answers during a two-way dialogue and receive targeted feedback from NLP techniques interpreting the spoken or written words. This new interaction for learning factual knowledge may be much more motivating and engaging, and may also be more effective at providing adaptive feedback and promoting deeper learning [11].

Given this potential for conversational approaches to enhance learning, we designed and built QuizBot, a dialogue-based adaptive learning system for students to learn and memorize factual knowledge in science, safety, and advanced English vocabulary. These three subjects were chosen because they cover diverse topics in medicine, language, and rules. They can represent important subclasses of factual knowledge that are usually learned outside the classroom setting.

On the technical side, QuizBot leverages the supervised Smooth Inverse Frequency (SIF) algorithm [2] for automatic answer grading and the DASH model [39] for adaptive question sequencing. On the design side, we created *Frosty*, an encouraging tutoring agent that provides targeted feedback to learners based on their inputs (see Figure 2). The design of QuizBot was inspired by previous studies [9, 13, 20] to leverage the *persona effect*, the strong positive impact of animated agents on learning experience [38].

To determine the impact of QuizBot on learning, we evaluated it against a carefully designed flashcard app, the typical medium for learning factual knowledge, through two controlled within-subject studies. We aimed to closely match the flashcard app to QuizBot in order to target assessment at the impacts of the conversational components. Specifically, the flashcard app used the same DASH algorithm for adaptive question selection, and a single pool of questions and answers was subdivided for the flashcard app and QuizBot.

In the first within-subject study with 40 students, when the number of practice items was held constant for both flashcards and QuizBot, students scored substantially better on recall (fill-in-the-blank) and recognition (multiple-choice) with QuizBot than for items trained using flashcards (66.3% vs. 45.2% for recall and 87.2% vs. 65.8% for recognition). However, the time taken was longer with QuizBot than flashcards. In the second within-subject study with 36 students, we allowed learners to voluntarily allocate their time between the two apps. We found students spent 2.6x more time on QuizBot, and that students performed equivalently on recognizing items but significantly better with QuizBot at recall (with an effect size of .45). These results suggest that QuizBot is more engaging to use and more effective at recall and equally effective at recognition in typical user-driven scenarios. In normal use, QuizBot may be less efficient per unit time, but still yields improved learning on recall due to users voluntarily choosing to use it substantially more.

This work has three chief contributions. First, QuizBot is the first chat-based learning system for factual knowledge memorization outside of classroom settings. Moreover, we show its effectiveness and engagement through rigorous comparison studies with a traditional learning tool for knowledge memorization, and our results demonstrate benefits of using chatbots to learn factual knowledge, especially for casual learning. Lastly, our results also reveal inefficiencies of chat-based learning systems, and we offer design suggestions for building improved future educational chatbot systems.

2 RELATED WORK

Our work was built upon previous studies on natural language tutoring systems, semantic similarity algorithms, and memory models.

Conversational Systems in Education

While the term *chatbot* is used in various contexts, the NLP community traditionally uses the phrase *conversational agent* for conversational systems and leaves chatbot for a subset of systems mostly handling casual conversation [33]. In our

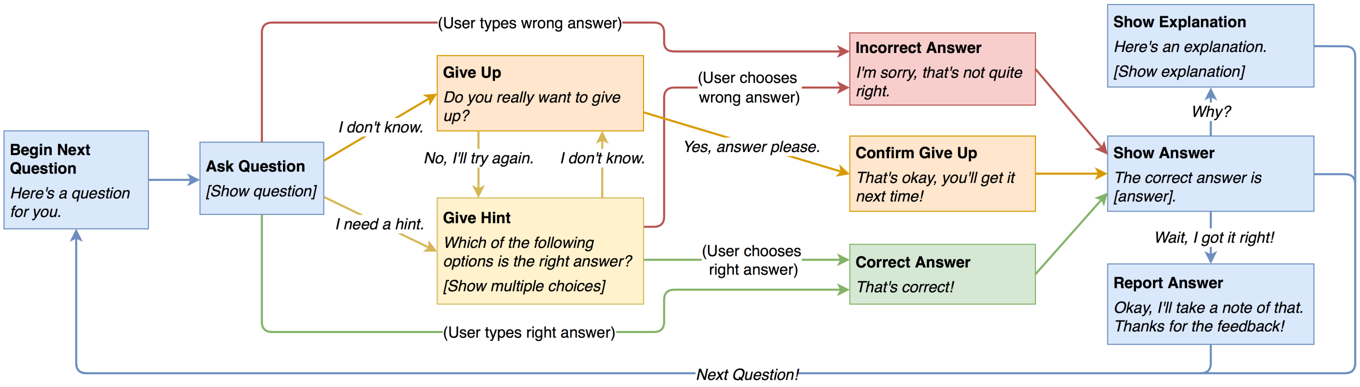

(User types wrong answer)
Begin Next Question
Here's a question for you.
Ask Question
[Show question]
I don't know.
I need a hint.
Give Up
Do you really want to give up?
No, I'll try again.
I don't know.
Give Hint
Which of the following options is the right answer?
[Show multiple choices]
(User chooses wrong answer)
Yes, answer please.
(User chooses right answer)
(User types right answer)
Incorrect Answer
I'm sorry, that's not quite right.
Confirm Give Up
That's okay, you'll get it next time!
Correct Answer
That's correct!
Show Explanation
Here's an explanation.
[Show explanation]
Why?
Show Answer
The correct answer is [answer].
Wait, I got it right!
Report Answer
Okay, I'll take a note of that. Thanks for the feedback!
Next Question!

Figure 2: The conversational state machine of QuizBot including typical sample responses.

work, we use "chatbots" and named our educational conversational agent *QuizBot*, as it was designed as a casual learning tool for a general audience to use on mobile devices.

Much prior work has built dialogue-based educational systems and evaluated their impacts on learning — for example, PACT Geometry Tutor for geometry problem-solving skills [1], Sofia for college math education [36], and Emile for social theories [41]. Interestingly, researchers have found inconsistent results regarding the effects of conversational systems on learning.

AutoTutor is a dialogue-based intelligent tutoring system teaching university students computer literacy and physics [18, 19] based on constructivist learning theories [16]. Laboratory and classroom studies show it led to learning gains of nearly one letter grade [17]. Ms. Lindquist is another natural language tutoring system that teaches high school students algebra [24]. Results from classroom use of Ms. Lindquist show that conversation had strong positive impacts on student motivation. Although students practiced fewer problems, they learned more per problem by being more engaged in a dialogue. Its authors characterized this type of systems as "less is more" [23]. Piagetbot, a conversation-based educational application emulating Jean Piaget, was created to teach Piagetian knowledge. Evaluation reveals that Piagetbot was less preferred by students and led to worse learning outcomes in comparison to a text-based interface [25].

Despite different effects on learning, all of these conversational tutoring systems were designed as formal study tools in traditional classroom settings. Altogether, this makes for a lack of research on educational chatbots for memorizing factual knowledge in a casual learning setting, as well as systematic studies evaluating their impacts on learning.

Semantic Similarity Algorithms

Semantic understanding of students' natural language inputs is critical to building educational chatbots. Recently, various general sentence similarity papers have advanced the field. The Universal Sentence Encoder [10] is a sentence level embedding method independent of existing word embeddings. The Smooth Inverse Frequency algorithm [2] trains models based on preexisting word embeddings like GloVe [45], weighting each word in the sentence according to inverse frequency. A number of papers also discuss the possibility of automating short answer grading using techniques such as graph alignment techniques [40], grading constraints [51], and inductive logic programming [53].

We decided to choose general sentence similarity models because they do not rely heavily on labeled data, are easy to generalize, and can incorporate out-of-vocabulary words.

Memory Models

One of the simplest and oldest models of human memory is the Ebbinghaus' exponential forgetting curve [14], which states that memory retention decreases at an exponential rate over time. Continued studies formalized *the spacing effect*, the observation that people tend to remember things more effectively if they use spaced repetition practice (short study periods spread out over time) as opposed to massed practice (i.e., "cramming") [50].

Physical tools like flashcards that take advantage of the spacing effect have existed for centuries. Recently, electronic spaced repetition software like SuperMemo [55] and Anki [15] has been developed to take advantage of automation of spaced repetition and increased flexibility to further optimize human learning. Although more customizable, these applications rely on rules preset by the SuperMemo algorithm [55] to schedule reviews. In [46], a number of algorithms were compared on various different memory models to show the efficacy of different strategies, including SuperMemo [55] and a threshold policy that selects an item with the likelihood closest to a chosen fixed threshold [5].

Recent work in cognitive psychology has led to the development of a new memory model, coined DASH [39], which combines aspects of power-law forgetting curves [54] with personalized features such as item difficulty, student ability, and study history.

3 DESIGN OF LEARNING SYSTEMS

We first describe QuizBot and its three key components: a dialogue system, a semantic similarity model, and an adaptive question sequencing algorithm. We then present the flashcard app used as a comparison system in our evaluation.

QuizBot System

QuizBot consists of two modes: a state-machine based quiz mode and a casual chat mode. The quiz mode is based on a rule-based chat system combined with a supervised sentence semantic similarity model.

Figure 2 illustrates one round of quiz mode interaction between a user and QuizBot. In the quiz mode, QuizBot asks a user a question selected by our question sequencing algorithm. A user then has three options: type in the answer if they know it, tap on the “Hint” button, or tap on the “I don’t know” button. If a user types and sends their answer to QuizBot, the chatbot will evaluate the correctness of the response by using an answer similarity computation algorithm. The model will return the cosine similarity between the correct answer and the user’s. Based on our empirical evaluation, QuizBot uses a threshold of 0.9 to decide if the user’s answer is correct and then passes the binary response to the spaced repetition model for selecting the next question. If the user asks for a hint, the chatbot will present the correct answer together with a list of distractors. The user can respond by tapping on any of the choices presented. After the user sees the correct answer, they can tap on the “Why” button for a short explanation. The interactions between the user and QuizBot are mixed between both typing and button selections; while inputting an answer is typing based, selecting from multiple choices and asking for an explanation are button based. The reasoning behind this mixed modality is to ensure both flexibility and efficiency regarding user interactions with QuizBot. The casual chat mode is user-initiated and rule-based; users can ask QuizBot for jokes or fun facts to take a break from the primary learning activity.

To evoke the same feelings one might have when having a conversation with another person, we took several steps to incorporate real-world conversational elements into QuizBot’s design. For example, we provided a wide variety of different responses to common conversation states and implemented mechanisms that enable QuizBot to provide hints and explanations when the user asks for them as well as positive reinforcement feedback that is typical of a study partner. In terms of conversational aspects common to chat-based interfaces, we incorporated graphical images and emojis in an attempt to replicate human-like behavior.

One other major design decision we made was to use an embodied conversational agent (electronic agents visually presented in the computer interface with some kind of human, animal, or fantasy form [4]) by personifying QuizBot as a penguin named *Frosty*, presented as a friend who is there to help the user learn. Prior work shows the presence of such visual imagery can improve performance on learning and memory tasks [4, 9, 42]. Figure 1 shows a conversation between a user and Frosty.

Semantic Similarity Between Short Sentences

We investigated various metrics to measure sentence similarity between user input and actual answers. Tensorflow included a package for a universal sentence encoder [10], which we compared to various *smooth inverse frequency* (SIF) models [2]. SIF implementations require a word embedding to identify a similarity metric for words. We decided to use GloVe [45], as we needed a general global word embedding. However, given much of our domain included vocabulary that was out of GloVe’s scope (e.g., particular scientific words in our science corpus [32]), we needed a way to extend GloVe without retraining a custom set. For this, we used Mittens [12] to utilize a domain-specific corpus to map new words with similar word embeddings to existing GloVe embeddings.

The unsupervised SIF implementation [2] maps s, the word vector embedding to v_s, a new vector embedding:

$$v_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w$$

where w is each word in s, v_w is the matching word embedding, $p(w)$ is the global word frequency of the word, and a is a small correction factor (10^{-3} in our case).

The SIF model additionally forms a matrix from all sentence embeddings and takes the first singular vector u to remove that singular component from each sentence. We fit our sentence model to the real-answer sentences in our dataset (we discuss how we collected it in Section 4) to get this singular vector, and then the transformation $v_s = v_s - uu^\top v_s$ is applied to all other sentences.

The supervised version of SIF took in respective similarity score data of pairs of sentences ranked from 1 to 5 and trained a neural network on the sentence similarity scores [2].

To compare these different models, we utilized a heatmap representation that included similar words next to each other and saw how different models performed on a set of test sentences collected from pilot studies (Figure 3). As evidenced by a higher concentration of red on the diagonal and greater white on the outside, the supervised SIF model on the right performed the best. Additionally, the supervised model allows for more flexibility and additional training data as we collect more data from pilot user studies. The final model we used in user studies was trained on 47 pilot users’ data.

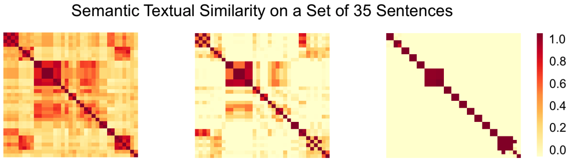


Figure 3: Heatmaps for different sentence embeddings: (left) Tensorflow's universal sentence encoding, (middle) unsupervised SIF, (right) supervised SIF with labeled data collected from pilot studies.

Question Selection Model

Our question selection model utilized a random model for sufficient question exploration with a review scheduled for every 5 questions chosen through a threshold policy based on the DASH model [39]. The DASH model was chosen because it includes various parameters that can be captured through collaborative filtering and can provide feedback for the mastery of different items.

The DASH model we adopted is mathematically formalized through a series of equations:

$$Pr(R_{si} = 1) = m(1 + h \cdot T)^{-f}$$

where R_{si} is a variable representing the retention rate of a particular item by a particular student. $h \in \mathbb{R}_+$ is the decay constant and $T \in \mathbb{R}_+$ represents the time elapsed since last review.

$$m = \sigma(a - d + h_\theta(\mathbf{t}_{1:k}, \mathbf{z}_{1:k-1}))$$

where σ is the logistic sigmoid function, a represents the student ability, d represents the particular item difficulty, and $\mathbf{t}_{1:k}, \mathbf{z}_{1:k-1}$ represents the outcomes of the previous k reviews of the item.

$$h_\theta(\mathbf{t}_{1:k}, \mathbf{z}_{1:k-1}) = \sum_{w=1}^{W} \theta_{2w-1} \log(1 + c_w) + \theta_{2w} \log(1 + n_w)$$

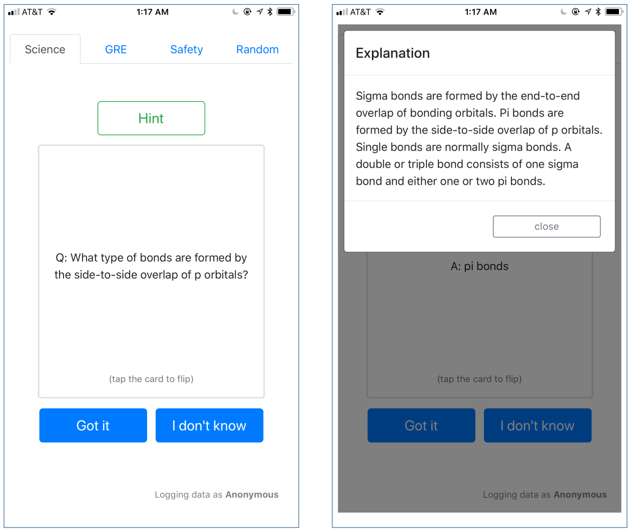
where w is an index over the time windows, c_w represents the number of times in the window that the student recalled the item, and n_w is the number of times in the window that the item was attempted. This function encapsulates student history with the item. θ are the window-specific weights, normally scaled based on time.

$$f = exp(\tilde{a} - \tilde{d})$$

where $\tilde{a}$ and $\tilde{d}$ represent additional student ability and item difficulty parameters.

Initialization of the latent parameters were based off of the parameters stated in [46], setting $a = \tilde{a} = 0$, $d \sim \mathcal{N}(1, 1)$, $\log \tilde{d} \sim \mathcal{N}(1, 1)$, $\log h \sim \mathcal{N}(0, 0.01)$, $\theta_{2w} = \theta_{2w-1} = 1/\sqrt{W - w + 1}$ and threshold initialized at 0.01.

Upon collecting information for our dataset from 65 Mechanical Turkers (details given in Section 4), item difficulties were recalculated to provide more accurate evaluations of item likelihoods by normalizing the correctness rates of the questions to have a mean of 1 and standard deviation of 1.



Q: What type of bonds are formed by the side-to-side overlap of p orbitals?

(tap the card to flip)

Logging data as **Anonymous**

Sigma bonds are formed by the end-to-end overlap of bonding orbitals. Pi bonds are formed by the side-to-side overlap of p orbitals. Single bonds are normally sigma bonds. A double or triple bond consists of one sigma bond and either one or two pi bonds.

A: pi bonds

(tap the card to flip)

Logging data as **Anonymous**

Figure 4: Screenshots of the flashcard app: (left) front side of the app where a question is presented, (right) back side of the app where an explanation is presented.

Flashcard System

To evaluate QuizBot fairly, we compared it against a flashcard app that we also implemented. Implementing our own flashcard app specifically allowed us to 1) incorporate features of popular flashcard apps currently available; 2) customize the question pool and question sequencing algorithms; 3) carefully control differences and similarities between the flashcard app and QuizBot; 4) log all the user interaction data into our own database.

To design the most intuitive interface for our flashcard app, we made a concerted effort to make it look and feel like the many other flashcard apps that are widely used in the educational space. In particular, we drew heavy inspiration from web and mobile-based flashcard apps like Quizlet [29], StudyBlue [30], and Magoosh [28]. These existing flashcard apps are popular among students, and all of them use the same tap-to-flip interaction paradigm. We decided to take several aspects from each app and integrate them into our own. For example, both Quizlet and StudyBlue utilize simple, functional designs that minimize possible external distracting factors. In addition, StudyBlue, as well as Magoosh, include self-evaluation metrics that allow users to mark which cards they have mastered. We implemented all these aspects, as well as others, into our flashcard app.

Our flashcard app can be seen in Figure 4. The app features a tabbed interface with a large flashcard in the middle and several buttons. The flashcard initially displays a question, and the user can tap it to "flip" the card and reveal the answer. Depending on the side of the card, either a hint button or explanation button is accessible allowing the user to view a list of possible answer choices or a short justification for the correct answer respectively. Users can tap the "Got it" or "I don't know" buttons below the flashcard to continue to a new question, and our spaced repetition algorithm uses the response to sequence questions for individuals. The tabs located at the top of the app give the user the ability to choose between subjects, as well as a "Random" option that mixes questions from all three.

To keep QuizBot and the flashcard app relatively consistent with one another, there are many features that are shared between them. First, the question pool is the same across both apps, which means all questions and answers come from the same source. The hint and explanation buttons in the flashcard app correspond respectively to the actions the user would take when asking for help or asking for why a particular answer is correct in Quizbot. To ensure that the volume of content consumption is about the same across both apps, we used the same question selection model and daily reminder notification system for both.

4 EXPERIMENT SETUP

In this section, we describe the experimental setup for our two user studies. Both studies were conducted remotely.

Participants

While building QuizBot, we iterated on the design with 47 university students and used their data to train the supervised SIF model, tune the DASH model, and improve the conversational and graphical design of QuizBot. Next, we launched QuizBot and recruited 80 college students and alumni through fliers, social networks, and mailing lists. Based on the order they were recruited, 40 of them participated in the first within-subject study that controlled the number of repetitions, and 40 of them (4 dropped out) participated in the second within-subject study that evaluated the engagement levels of the two apps. The 76 students who finished the studies came from 12 different universities and over 20 different majors including computer science, mathematics, biology, history, communication, psychology, and more. Study 1 users were compensated $75 and Study 2 users were compensated $50 for their participation.

Apparatus

All participants used their own mobile phones to install our apps for the study. 51 were iPhones and 25 were Android phones. The back end of our QuizBot system was implemented in Python Flask, and the front end used Facebook Messenger's chat interface, which provides a platform to easily serve chatbots. The complementary flashcard app was implemented as a web app and converted to iOS and Android apps using Apache Cordova. We logged all the users' pertinent behaviors such as conversation logs, button clicks, and corresponding timestamps to a MySQL database for detailed data analysis at a later time.

Question Pool

Our question pool contained factual questions in the areas of science, safety, and advanced English vocabulary. We chose three different subjects for the system's question pool because we wanted to study how capable the system was in helping students learn factual knowledge in different areas. Every question has three parts: a question description, a correct fact-based answer, a set of 2 to 4 distractors (M=3.2, SD=0.8), and an explanation paragraph. On average, questions contained 18.2 words (SD=10.2) and correct answers contained 2.3 words (SD=2.6).

All the questions, distractors, and explanations were drawn from free online resources. Science questions were adapted from the *SciQ* dataset [32]. Safety questions were adapted from *MySafetySign*, a safety quiz website [43]. Verbal reasoning questions were adapted from *KMF*, a free GRE prep website [35]. Initially, we randomly selected 100 questions for each subject. Then four different researchers from our team went through each question and discarded overly easy or difficult questions. Lastly, we refined the selected questions to ensure their suitability for our learning systems.

Balancing Question Difficulty

It was important to make sure that the questions in our question pool had similar difficulty levels. After manually filtering the questions, we put them on Mechanical Turk and recruited 65 crowd workers to answer them. We required all the workers to have a bachelor's degree and at least a 98% HIT approval rate with at least 1000 completed HITs. We also added attention check questions to the online multiple choice quiz to ensure crowd workers answered the quiz in good faith.

For every question q, we assigned it a difficulty rating d computed as $d(q)$ = number of workers who answered q correctly / total number of workers. For a set of questions S, its *set difficulty* $d(S)$ was computed as the average of the difficulty ratings of all the questions in this set. We used dynamic programming to select two sets of an equal number of questions that had the smallest difference of difficulty, based on difficulty ratings collected from Mechanical Turk. In our studies, the set difficulty differences between all sets we used were below 0.1.

In addition, the difficulty parameters were fit into the DASH model to provide better question sequencing as discussed in Section 3.

5 EVALUATION AND RESULTS

Effectiveness and engagement are the two most important metrics in our evaluation of QuizBot. In particular, we proposed the following research questions:

(1) How engaging is QuizBot to learners in comparison to flashcards?
(2) How effective is QuizBot with helping learners in recognition and recall, both per number of practice items and per time spent, compared with flashcards?
(3) Given voluntary usage, which system is more effective on recognition and recall?

We performed two within-subject user studies to answer these research questions: Study 1 answered (2) and Study 2 answered (1) and (3). The following commonly adopted metrics were used to report our results:

Total repetitions: the total number of repetitions completed by learners. One question can be repeated many times.

Total questions: the total number of unique questions completed by learners.

Overall recognition / recall accuracy: the percentage of questions missed in the pre-test that were correctly answered on the multiple choice / fill-in-the-blank post-test, regardless of whether they are practiced during system use or not.

Practiced recognition / recall accuracy: the percentage of questions missed in the pre-test, introduced by the system, and were correctly answered on the multiple choice / fill-in-the-blank post-test.

Usage time: the total amount of time learners spent interacting with the system in the entire learning period. A threshold of 30 seconds was used to to decide if a learner was idle during any given study session.

Return rate: the total number of times learners returned to use the system in the entire learning period.

All results are presented using the sample mean (standard deviation) notation, and error bars on bar charts represent +/- 1 standard error. We use † to indicate a statistical difference at the .05 significance level in the measure of QuizBot against that of the flashcard app in the within-subject study.

Study 1: Evaluation of Effectiveness

The first experiment we ran was a within-subject study designed to evaluate the effectiveness of the two learning apps. We recruited 40 university students and alumni from 11 different universities. 23 were females and 17 were males. Their average age was 23.5 (*SD*=3.5).

The question pool of each app contained 48 questions, 16 questions for each subject. Questions in the QuizBot app and the flashcard app were distinct and their difficulty levels were carefully balanced as described previously. Since we wanted to evaluate how effective each app was with helping people memorize factual knowledge, we fixed the number of repetitions participants could perform for each question. Participants were required to use both apps to learn factual knowledge everyday in a five-day study period: 20 questions within each app in the first four days, and 16 questions in the last day, so that every question was practiced by every user exactly twice. Each app sequenced 48 questions using the same algorithm.

Participants were given a pre-test consisting of all 96 questions from the two question pools in the multiple-choice format, randomly mixed together. After the five-day study period, they were given two post-tests that consisted of the same questions, but one requiring fill-in-the-blank and another using multiple-choice. Post-tests were customized for each participant. Only questions participants answered incorrectly in the pre-test were included in the post-test. By asking learners to only answer questions they did not know before, we eliminated the effects of people's different prior knowledge on their final learning outcome. The average number of questions participants answered correctly in the pre-test of the flashcard questions pool was 19 (*SD*=6) and in the QuizBot question pool was 21 (*SD*=6). Thus, the average length of one post-test was 56 (*SD*=11). All of the fill-in-the-blank questions were presented to learners prior to the multiple-choice questions to avoid carryover effects. After collecting learners' answers, we manually graded their answers to the fill-in-the-blank questions.

To counterbalance our experiment, half of the participants learned with QuizBot first followed by the flashcard app every day, and the other half used the two apps every day in the opposite order. The order of which app to use first was randomly assigned and remained the same across all five days. Our analysis of variance shows that order did not exhibit any significant effect for any of our measures.

As shown in Figure 5, after practicing each question twice, learners correctly answered 87.2% (*SD*=10.3) of the multiple-choice questions and 66.3% (*SD*=20.9) of the fill-in-the-blank questions that they did not know in the pre-test with QuizBot, versus 65.8% (*SD*=21.0) and 45.2% (*SD*=23.4) with the flashcard app. The recognition accuracy differences of the two apps followed the normality assumption according to the Shapiro-Wilk normality test result (p=.62). A paired two-sample t-test was performed and the difference was significant for recognition accuracy (t_{39}=-6.94, p<.0001). Effect size was 1.10 with a power of 1.00. The differences in recall accuracy did not satisfy the normality assumption (p<.01 from the normality test). Therefore, we used a nonparametric paired two-sample Wilcoxon test. Results show that recall accuracy using QuizBot was statistically significantly higher than the recall accuracy using the flashcard app (Z=34, p<.0001). Effect size was 1.03 with a power of 1.00.

We also tested students' retained knowledge one week and two months after the study. The delayed post-tests were

Table 1: Results of Studies 1 and 2. Accuracies are in percent. F = Flashcard, Q = QuizBot. Subscripts indicate studies.

	Usage time (*m*)	Per question time (*s*)	Return rate	Total reps.	Total questions	Overall recall acc.	Practiced recall acc.	Overall recog. acc.	Practiced recog. acc.	One-week recog. acc.	Two-month recog. acc.
F_1	14.3 (7.4)†	8.9 (4.6)†	5 (0)	96 (0)	48 (0)	45.2 (23.4)†	45.2 (23.4)†	65.8 (21.0)†	65.8 (21.0)†	64.4 (20.7)†	43.9 (28.2)†
Q_1	46.0 (6.3)†	28.7 (4.0)†	5 (0)	96 (0)	48 (0)	66.3 (20.9)†	66.3 (20.9)†	87.2 (10.3)†	87.2 (10.3)†	83.6 (11.7)†	52.0 (31.2)†
F_2	14.4 (13.2)†	9.9 (7.2)†	8 (6)†	181 (285)	39 (31)	38.7 (25.5)†	60.1 (26.9)†	62.4 (25.9)	83.1 (21.2)†	-	-
Q_2	38.1 (16.9)†	37.2 (13.9)†	24 (7)†	69 (36)	30 (12)	50.8 (23.1)†	75.1 (25.0)†	67.3 (27.0)	94.1 (13.2)†	-	-

Table 2: Per-subject results of Study 1. Pre-test scores are out of 16 for all subjects. Accuracies are in percent. F = Flashcard, Q = QuizBot.

		Science	Safety	Vocabulary
Pre-test score	F	6.2 (2.5)†	5.5 (2.3)†	7.6 (3.4)
	Q	7.1 (3.3)†	6.6 (1.9)†	7.9 (3.5)
Recall accuracy	F	49.0 (28.8)†	50.1 (25.8)†	37.0 (28.7)†
	Q	70.0 (26.6)†	71.2 (21.9)†	60.4 (29.0)†
Recog. accuracy	F	68.7 (23.2)†	66.8 (23.2)†	61.8 (31.2)†
	Q	90.7 (12.7)†	90.0 (13.0)†	81.7 (18.8)†
One-week recog. acc.	F	68.3 (23.6)†	65.2 (23.7)†	58.6 (30.6)†
	Q	90.3 (12.7)†	84.8 (13.4)†	78.3 (21.2)†
Two-month recog. acc.	F	47.7 (31.7)†	42.1 (27.7)†	39.7 (30.3)†
	Q	58.9 (34.7)†	51.5 (30.0)†	45.9 (38.7)†

identical to the immediate multiple-choice post-test except for the randomized question order. All 40 users responded to the one-week post-test. Results in Table 1 show that students' knowledge on the initial and delayed post-test was almost identical (p = .29 for the flashcard app and p=.12 for QuizBot) after one week. 39 out of 40 users responded to the two-month delayed post-test. People's recognition accuracy after two months was 52.0% (SD=31.2%) for QuizBot and 43.9% (SD=28.2%) for the flashcard app. The difference was statistically significant with a p-value of .006 and an effect size of .47. Therefore, after two months, there is a smaller but still significant beneficial effect on recognition from using QuizBot over flashcards. These results demonstrate that given a fixed number of items practiced, QuizBot was significantly more effective in helping people memorize factual knowledge.

Looking at subjects separately, students' performance is shown in Table 2. Although people had different prior knowledge in the three subjects, they improved more with QuizBot than with flashcards, on both recognition and recall. The improvement was statistically significant for all three subjects and for both recall and recognition. This shows that the effectiveness of QuizBot may well generalize to different domains.

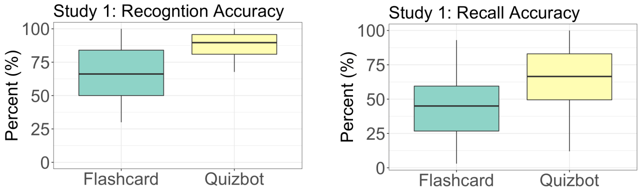


Figure 5: Study 1 overall (practiced) recognition (left) and recall (right) accuracy.

We then examined the time taken to finish all of the 96 repetitions with each of the two apps. While it took on average 28.7*s* (*SD*=4.0) to practice one question with QuizBot, it took 8.9*s* (*SD*=4.6) to practice one question with the flashcard app. The difference was statistically significant: t_{39}=-31.6, p<.0001 with an effect size of 5.0. Although time per question is longer, users rated QuizBot higher in the User Engagement Scale (short form) [44] and the difference was statistically significant (t_{39}=-2.90, p<.01) with an effect size of .46, based on a paired two-sample t-test. Users also preferred QuizBot strongly for casual learning as shown in Figure 6 (left). Our second within-subject study evaluates student motivation and performance in a casual learning setting.

Study 2: Evaluation of Engagement

The second within-subject study was designed to evaluate learner engagement level and performance given voluntary usage time. We recruited 36 users (15 female, 20 male, and 1 chose not to say) from 8 different universities in this experiment. The average age of these 36 users was 22.8 (*SD*=3.1). Every user was required to use both apps everyday over a five day period. The QuizBot app and the flashcard app had the same non-overlapping set of 48 questions, as in the first within-subject experiment. Users chose the amount of time to spend on each app of their own volition, and they were advised to spend about 10 minutes on each of the two apps in total to ensure some basic knowledge gain. We used DASH as the question sequencing algorithm in this experiment.

The 36 participants' total usage time for both apps is shown in Table 1. As can be seen, users spent on average 38.1 minutes (*SD*=16.9) learning with the QuizBot app while only 14.4 minutes (*SD*=13.2) with the flashcard app. A paired two-sample t-test shows that the usage time differences were statistically different: t_{35}=-6.08, p<.0001. The effect size of the difference was 1.01 with a power of 1.00. Users' return rate with QuizBot was also higher and the difference was statistically significant: t_{35}=-12.4, p<.0001. The effect size was 2.06 with a power of 1.00. The significantly longer self-driven usage time and higher return rate on QuizBot demonstrates its engagement compared to the flashcard app.

All participants were given 54 pre-test multiple choice questions with 27 questions randomly selected from each app's question pool, or 9 questions from each subject. For the post-test, like Study 1, participants were tested on questions they answered incorrectly in the pre-test in a fill-in-the-blank format followed by a multiple choice format. Participants on average got 11 flashcard questions (*SD*=4) and 12 QuizBot questions (*SD*=5) correct in the pre-test, so their post-test consisted of on average 16 flashcard questions (*SD*=4) and 15 QuizBot questions (*SD*=5). The total number of repetitions users did with each app is shown in Table 1. As can be seen, users had a higher number of repetitions and unique questions completed with the flashcard app. However, the results in Table 1 show that users performed better on QuizBot questions in the post-test, both on all the post-test questions (overall) and on questions seen within each app (practiced). Paired t-tests show that the performance difference was statistically significant in practiced recognition accuracy (p<.05 with an effect size of .49), overall recall accuracy (p<.01 with an effect size of .45), and practiced recall accuracy (p<.05 with an effect size of .47), and not in overall recognition accuracy (p=.42). This shows that people voluntarily spent enough time on QuizBot such that their recognition performance was similar and their recall performance was significantly better than when using the flashcard app.

Subjective Rating and Qualitative Feedback

Users' preferences between the two apps are shown in Figure 6 and their subjective ratings across four categories are shown in Figure 7. As can be seen, more than 68% of users in both experiments liked QuizBot and more than 63% preferred using it for casual learning, but the percentage dropped when it came to preparation for exams, likely due to its lower efficiency. QuizBot was rated as significantly more fun (p<.01 for Study 1 and p<.0001 for Study 2) and more effective (p<.05 for Study 1 and p<.001 for Study 2) than the flashcard app. QuizBot was rated on par with the flashcard app in terms of ease of use (p=.06 for Study 1 and p=.90 for Study 2) and efficiency (p = .29 for Study 1 and p=.81 for Study 2).

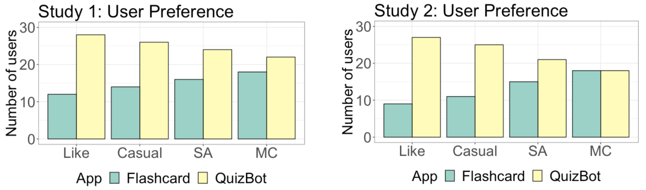


Figure 6: Preferences of users from Study 1 (left) and Study 2 (right). From left to right: which app users like better (Like) / prefer to use for casual learning (Casual) / for short-answer exams (SA) / for multiple-choice exams (MC).

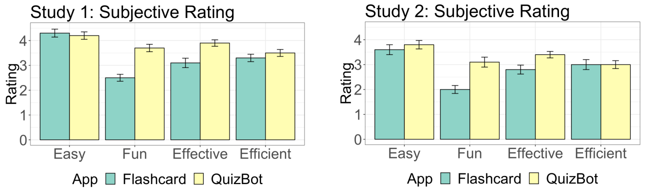


Figure 7: Study 1 (left) and Study 2 (right) users' subjective ratings on a 1 to 5 scale. Higher is better.

We surveyed all participants on their opinions about QuizBot. Their general attitude was that QuizBot was more interactive and more engaging, though less efficient, as illustrated by the following representative responses:

On communication and interactivity: *"I could actually answer and type it out rather than just looking at flashcards."*

On engagement and feedback: *"QuizBot was more interactive and engaging. The feedback (positive or negative) was also motivating."*

On graphics: *"The penguin graphics are really cute."*

On conversations: *"The penguin had a lot of personality, and I enjoyed trying to trick the chatbot."*

On speed: *"I can go over questions and answers very fast. Sometimes I have to wait for QuizBot to respond."*

On context of use: *"It is something that I would definitely use if available and can easily be used during a bus ride etc."*

6 DISCUSSION

In this section, we discuss and answer our research questions based on evaluation results, and we offer suggestions for designing further improved chat-based learning systems.

NLP Accuracy and Errors

Our preliminary analysis suggests QuizBot rarely graded answers incorrectly. To assess this, we randomly selected 11000 (out of 43,956) consecutive conversational logs from Study 1 and manually examined users' utterances graded by our system. These logs contained 1052 questions and 144 of them were answered by users via typing. 139 questions were correctly understood and graded by our algorithm, leading to an accuracy rate of 96.5%. Of the remaining 5 incorrectly graded questions, 1 was because the system could not handle a typo (a user typed “ture” instead of “true”); 3 were because our algorithm penalized short answers and users tended to type terse answers (e.g., a user typed “class b” while the correct answer was “class b carbon dioxide fire extinguisher”); the remaining was because the system was not perfect at understanding phrases (a user typed “try breast compression only” and the system regarded it as correct while the correct answer was “perform chest compressions only”).

Also, we provided a “Report Bug” button in QuizBot so that users could report incorrectly graded natural language answers and we could update the model in real time. In our experiments, less than 1% of answers were reported by users as incorrectly graded. High system grading accuracy may therefore be an important contributor to QuizBot’s generally positive user experience.

Effectiveness

Our evaluation demonstrated that when the number of practice items is fixed, QuizBot is more effective in boosting learning in terms of both recognition and recall. We believe QuizBot’s effectiveness may be attributed to the following four factors.

First, the testing effect through active recognition and recall in QuizBot is more effective than the passive memorization that is typical in flashcard software. As multiple studies have shown, active retrieval via attempts to recollect memory produces better retention than only restudying the information [31, 47, 48].

Second, feedback helps enhance the testing effect [3, 37], regardless of whether the attempted answers are correct or not [6]. With QuizBot’s underlying semantic similarity algorithm, students were able to get targeted feedback on both their typed answers and multiple choice answers.

Third, writing out knowledge may improve its mastery compared to recognizing an answer from provided options. In Study 1, students correctly recognized 95.5% (*SD*=8.6) and recalled 44.3% (*SD*=26.6) of QuizBot’s post-test questions practiced by typing, in comparison to correctly recognizing 88.5% (*SD*=11.6) and recalling 35.0% (*SD*=16.5) of questions practiced by selecting from multiple choices. In post-study questionnaires, 65.8% of users in Study 1 and 63.9% of users in Study 2 confirmed that typing was more effective for them compared to choosing from multiple choices.

Fourth, considering the persona effect identified by prior work [13, 20, 38], a carefully designed chat-based interface leads to better student motivation and ultimate learning outcomes.

Our second finding is that given voluntary usage, QuizBot led to significantly better learning outcomes on recall and similar outcomes on recognition, as seen in Study 2. These results are consistent with findings reported in the literature on educational agents [22–24]. Although students practiced fewer problems due to the relative inefficiency of conversational systems, students were more engaged in the problems they did and more willing to be tutored by the conversational tutoring systems. As a result, they performed equally well or even better, highlighting that sometimes “less is more” [23].

In addition, in study 2, students improved more using QuizBot on fill-in-the-blank than on multiple-choice post-test questions. Students in both studies also preferred QuizBot more for short answer tests than for multiple-choice tests as illustrated in Figure 6. This is consistent with what Heller and Procter hypothesized in their previous work [25]. Traditional ways for preparing for traditional multiple-choice tests are likely sufficient, and novel chat-based interfaces might be better for studying non-traditional types of questions such as fill-in-the-blank.

Trade-off Between Engagement and Efficiency

As hypothesized, our participants were more engaged with QuizBot. They spent 2.6x more time interacting with QuizBot in Study 1 and also rated it higher in both studies as shown in Figure 6 and 7. Many users commented that they preferred to learn using QuizBot because it was “fun, interactive, and felt like a real study partner.”

However, the fun conversational side of QuizBot also led to inefficiencies in learning. After analyzing the participants’ conversational logs and timestamps, we broke down the interaction time with QuizBot, shown in Figure 8. As can be seen, 11.7% of the total time was not spent on learning with QuizBot. For example, we deliberately added delays to QuizBot’s sequential conversations to make it feel like a real person was typing, which led to a 4.1% manual delay. Users also spent 2.1% of the total time just chatting with the chatbot for fun. Had we removed these casual aspects from QuizBot, users’ interest or motivation in learning might be negatively affected. Hence, there is a trade-off between engagement and efficiency.

Potential Use Cases for Applying Bots in Education

Our proof-of-concept work sought to explore how a chat-based agent might be beneficial for supporting fact-based learning for informal learning settings. To do so, we compared a chatbot using an animated conversation and a natural language targeted feedback mechanism with a flashcard app; the adaptive learning algorithm for selecting the next item was the same for both apps. Our results showed that after practicing equal numbers of items with the two apps, people better recalled and recognized items practiced with the chatbot, although the chatbot was much slower. However, as we

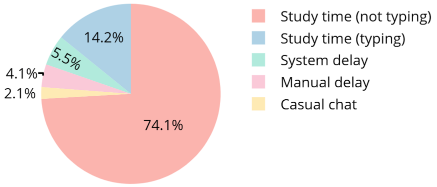


Figure 8: Time proportion of different user and system activities in QuizBot in our within-subject Study 1.

were curious about informal learning support, we also examined how less prescribed usage and learning outcomes varied between the conditions. Interestingly, even though the flashcard application was more efficient per time spent, people spent substantially longer using the chatbot app, and also ended up with substantially improved recall performance (and equal recognition) to the flashcard app. This suggests future implications for where and for whom chatbots may be useful in learning factual knowledge. In particular, for highly motivated students with limited time, flashcards may yield better outcomes; but for those either without a hard time bound or who are less motivated to learn this material, chatbots may yield stronger learning outcomes.

Limitations and Future Work

One obvious disadvantage of QuizBot is its inefficiency compared to the flashcard app. To better understand this, we analyzed the time proportion of different activities as shown in Figure 8. Some delays can be improved. For example, system delay time can be shortened if we have faster algorithms and better Internet connections (our machine learning models were computed on the cloud). Users' typing contributed to 14.2% of the total usage time. Previous research shows that speech-based input methods are about three times faster than keyboard-based input methods [49]. To improve users' input speed, we could therefore consider supporting speech in addition to keyboard based conversations.

Another direction that could be taken to improve QuizBot is to leverage richer natural-language input from learners to enhance spaced repetition models. We were able to collect high dimensional natural language inputs from learners. Currently, the system only assigns cosine similarity scores to natural language inputs and truncates them to binary inputs to fit into DASH. A more elaborate way to handle these high dimensional inputs will likely yield a better learning system.

Also, we compared QuizBot against a flashcard app that was designed based on current popular flashcard software. A natural option given our findings would be a flashcard system where learners have to type in the answer. This hybrid app may have the advantages of both QuizBot and flashcard apps: efficacy, engagement, and efficiency. It would be compelling to build and study such hybrid systems going forward.

Further, it is important yet challenging to identify the key aspects in a systems level contribution. While we did not have a study separating out the impact of each component, we interviewed our users on how QuizBot affected their engagement and learning. Based on the qualitative feedback received, it is our belief that the avatar and conversations contributed the most to improving engagement, and that the active recall and feedback helped most with strengthening memorization. However, understanding the precise contributions of each system feature remains an important issue for future work.

Lastly, we evaluated learners' longer-term learning outcomes but not their longitudinal engagement levels. As other researchers have studied, once the novelty effect of the animated agent has decayed, users' engagement level may drop [13]. We computed the over-time usage statistics within the study period to investigate this novelty effect. For QuizBot, the average usage time (in minutes) of all 36 Study 2 users from day 1 to day 5 was 7.6, 8.3, 7.6, 6.6, 8.1. For flashcards, the average usage time was 4.4, 3.3, 2.0, 3.2, 1.5. Therefore, when users were asked to use two apps voluntarily, they spent a similar amount of time day to day on QuizBot and a decreasing amount of time on the flashcard app. This indicates that novelty may not have be a primary factor contributing to the increased engagement and learning we saw during our five-day intervention, though further study is necessary to fully understand long term usage behaviors and attitudes.

7 CONCLUSION

In this research, we designed, built, and evaluated QuizBot, a conversation-based learning system for factual knowledge that incorporates some of the latest progress in semantic similarity models and spaced repetition algorithms. We compared QuizBot to a traditional flashcard learning tool in two user studies. In general, we found that QuizBot helped learners recognize 21.4% more (and recall 21.0% more) questions than the flashcard app. Despite taking longer to learn with QuizBot, learners were much more engaged with Quizbot, and this led to their improved performance on recall and similar performance on recognition. Our work also offers design suggestions to further improve chat-based learning systems. With AI algorithms becoming increasingly powerful, we hope our work will attract more researchers to develop effective and engaging natural language tutoring systems.[1]

ACKNOWLEDGMENTS

We would like to thank TAL Education Group for funding this research project and Anna Wang for designing the graphics and the conversational visual style of QuizBot.

[1] Our project site can be found at hci.stanford.edu/research/smartprimer

REFERENCES

[1] Vincent Aleven, Kenneth R Koedinger, and Karen Cross. 1999. Tutoring answer explanation fosters learning with understanding understanding. In *Proceedings of the 9th International Conference on Artificial Intelligence in Education*. 199–206.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *International Conference on Learning Representations* (2017).

[3] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of Educational Research* 61, 2 (1991), 213–238.

[4] Robbert-Jan Beun, Eveliene de Vos, and Cilia Witteman. 2003. Embodied Conversational Agents: Effects on Memory Performance and Anthropomorphisation. In *Intelligent Virtual Agents*, Thomas Rist, Ruth S. Aylett, Daniel Ballin, and Jeff Rickel (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 315–319.

[5] Robert A Bjork. 1994. Memory and metamemory considerations in the. *Metacognition: Knowing About Knowing* 185 (1994).

[6] Andrew C Butler, Jeffrey D Karpicke, and Henry L Roediger III. 2008. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 4 (2008), 918.

[7] Robert M Carini, George D Kuh, and Stephen P Klein. 2006. Student engagement and student learning: Testing the linkages. *Research in Higher Education* 47, 1 (2006), 1–32.

[8] Mark Carrier and Harold Pashler. 1992. The influence of retrieval on retention. *Memory & Cognition* 20, 6 (1992), 633–642.

[9] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied Conversational Agents*. MIT press.

[10] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *Computing Research Repository* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[11] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. https://doi.org/10.1080/00461520.2014.965823

[12] Nicholas Dingwall and Christopher Potts. 2018. Mittens: An Extension of GloVe for Learning Domain-Specialized Representations. *arXiv preprint arXiv:1803.09901* (2018).

[13] Kathryn Hershey Dirkin, Punya Mishra, and Ellen Altermatt. 2005. All or nothing: Levels of sociability of a pedagogical software agent and its impact on student perceptions and learning. *Journal of Educational Multimedia and Hypermedia* 14, 2 (2005), 113–127.

[14] Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of Neurosciences* 20, 4 (2013), 155.

[15] Damien Elmes. 2015. Anki. http://ankisrs.net

[16] Catherine Twomey Fosnot and Randall Stewart Perry. 1996. Constructivism: A psychological theory of learning. *Constructivism: Theory, Perspectives, and Practice* 2 (1996), 8–33.

[17] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

[18] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 180–192.

[19] Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Roger Kreuz, Tutoring Research Group, et al. 1999. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research* 1, 1 (1999), 35–51.

[20] Agneta Gulz and Magnus Haake. 2006. Design of Animated Pedagogical agents - A Look at Their Look. *International Journal of Human-Computer Studies* 64, 4 (April 2006), 322–339. https://doi.org/10.1016/j.ijhcs.2005.08.006

[21] Marti A Hearst. 2015. Can Natural Language Processing Become Natural Language Coaching?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 1245–1252.

[22] Neil T. Heffernan. 1998. Intelligent Tutoring Systems Have Forgotten the Tutor: Adding a Cognitive Model of Human Tutors. In *Conference on Human Factors in Computing Systems (CHI '98)*. ACM, New York, NY, USA, 50–51. https://doi.org/10.1145/286498.286524

[23] Neil T Heffernan. 2003. Web-based evaluations showing both cognitive and motivational benefits of the Ms. Lindquist tutor. In *Artificial Intelligence in Education*. 115–122.

[24] Neil T Heffernan and Kenneth R Koedinger. 2002. An intelligent tutoring system incorporating a model of an experienced human tutor. In *International Conference on Intelligent Tutoring Systems*. Springer, 596–608.

[25] Bob Heller and Mike Procter. 2007. Conversational Agents and Learning Outcomes: An Experimental Investigation. In *EdMedia: World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), 945–950.

[26] Paul Jen-Hwa Hu and Wendy Hui. 2012. Examining the role of learning engagement in technology-mediated learning and its effects on learning effectiveness and satisfaction. *Decision Support Systems* 53, 4 (2012), 782 – 792. https://doi.org/10.1016/j.dss.2012.05.014

[27] Shouping Hu, George D Kuh, and Shaoqing Li. 2008. The effects of engagement in inquiry-oriented activities on student learning and personal development. *Innovative Higher Education* 33, 2 (2008), 71–81.

[28] Magoosh Inc. 2018. Magoosh. https://magoosh.com/

[29] Quizlet Inc. 2018. Quizlet. https://quizlet.com/

[30] StudyBlue Inc. 2018. StudyBlue. https://www.studyblue.com/

[31] William James. 1890. *The Principles of Psychology*. Dover Publications.

[32] Nelson F. Liu Johannes Welbl and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Workshop on Noisy User-generated Text*. Copenhagen, Denmark.

[33] Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (1st ed.). Prentice Hall PTR.

[34] Jeffrey D Karpicke and Henry L Roediger. 2008. The critical importance of retrieval for learning. *Science* 319, 5865 (2008), 966–968.

[35] kmf.com. 2018. GRE KMF. https://gre.kmf.com/. Accessed: 2018-06-01.

[36] Oliver Knill, Johnny Carlsson, Andrew Chi, and Mark Lezama. 2004. An artificial intelligence experiment in college math education. *Preprint available at http://www. math. harvard. edu/˜ knill/preprints/sofia. pdf* (2004).

[37] Raymond W Kulhavy and William A Stock. 1989. Feedback in written instruction: The place of response certitude. *Educational Psychology Review* 1, 4 (1989), 279–308.

[38] James C. Lester, Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone, and Ravinder S. Bhogal. 1997. The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 359–366. https://doi.org/10.1145/258549.258797

[39] Robert V Lindsey and Michael C Mozer. 2016. Predicting and improving memory retention: Psychological theory matters in the big data era. In *Big Data in Cognitive Science*. Psychology Press, 43–73.

[40] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, 752–762. http://dl.acm.org/citation.cfm?id=2002472.2002568

[41] R Moore and G Gibbs. 2002. Emile: Using a chatbot conversation to enhance the learning of social theory. *Univ. of Huddersfield, Huddersfield, England* (2002).

[42] Roxana Moreno, Richard E Mayer, Hiller A Spires, and James C Lester. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction* 19, 2 (2001), 177–213.

[43] MySafetySign. 2018. Free Safety Quizzes - Workplace Safety Rules. https://www.mysafetysign.com/safety-quiz. Accessed: 2018-06-01.

[44] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28 – 39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[46] Siddharth Reddy, Sergey Levine, and Anca Dragan. 2017. Accelerating Human Learning with Deep Reinforcement Learning. *NIPS'17 Workshop: Teaching Machines, Robots, and Humans* (2017).

[47] Henry Roediger and Andrew Butler. 2010. The critical role of retrieval practice in long-term retention. 15 (10 2010), 20–7.

[48] Henry Roediger and Jeffrey Karpicke. 2006. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science* 1, 3 (2006), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x PMID: 26151629.

[49] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4, Article 159 (Jan. 2018), 23 pages. https://doi.org/10.1145/3161187

[50] Burr Settles and Brendan Meeder. 2016. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1848–1858. https://doi.org/10.18653/v1/P16-1174

[51] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1070–1075.

[52] Joel Tetreault, Jill Burstein, and Claudia Leacock. 2015. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

[53] Alistair Willis. 2015. Using nlp to support scalable assessment of short free text responses. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. 243–253.

[54] John T Wixted, Shana K Carpenter, et al. 2007. The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science* 18, 2 (2007), 133.

[55] PA Wozniak and Edward J Gorzelanczyk. 1994. Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis* 54 (1994), 59–59.