Human Learning Meets Representation Learning

Matthew Shu[★]
Yale University
matthew.shu@yale.edu

Shi Feng[★]
University of Maryland
shifeng@cs.umd.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

Effective human learning requires the instructor to properly organize the learning material. To do so, the instructor must understand the topic and difficulty of each item, the student's current level of comprehension, and how the student's knowledge is affected by time and interventions. Automated instruction software like Duolingo already employ machine learning to build student models from practice data [Settles and Meeder, 2016, Hunziker et al., 2019, Upadhyay et al., 2020]. However, these models use manually crafted features which are insufficient for understanding the semantic relationships between items and how they impact student memory. To address this shortcoming, we incorporate pretrained textual representations into a neural student model: this allows our model to infer more about items the student has not explicitly studied in the app. We evaluate our system (KAR^{3}L) in a user study on a web-based flashcard learning app. Results from this four-month period indicate that incorporating better representations for semantic relationships improves models of human learning.

1 Introduction

A good teacher understands what a student knows and what they are prepared to learn next [Vygotsky, 1980]. The characteristics of the learning material (i.e., the *items*), as well as the *semantic relationships* between item content (e.g., *Mansa Musa* and *Mali*), lead to variations in memory across and within students. Accurate evaluation of memory ability by a *student model* is therefore crucial for automated instruction [Mozer et al., 2019]. A student model estimates current student recall probability for each item (e.g., a flashcard) and predicts how time and practice affects these estimates.

Given the student model, the teacher must decide which information to present when. In automated instruction, this task is handled by a *teaching policy*, which takes in estimates from the student model and suggests when to present an item. The objective is to meet student goals under constraints of time and effort [Mozer et al., 2019, Tabibian et al., 2019].

Teachers should teach items of appropriate difficulty. Students knowledgeable about American geography will likely gain little from flashcards associating state shapes with names. The effectiveness of a teaching policy therefore relies on accurate student model predictions of recall probability, a proxy for item difficulty. Combining psychological theories and machine learning, existing student models can adjust recall probability predictions for items using manually-crafted features derived from practice history, e.g., average recall rate and time of last review [Settles and Meeder, 2016, Hunziker et al., 2019, Tabibian et al., 2019, Upadhyay et al., 2020].

However, existing student models mostly ignore item content. When a human teacher asks *"Who was the second assassinated president of the United States?"* and the student correctly answers *"James Garfield"*, that teacher can infer the student likely also knows that Abraham Lincoln was the first assassinated American president. Existing student models are limited in making these inferences from semantic relationships between items to adjust predicted recall probability. In other words, existing models are largely *content-naïve*.

[★]Equal contribution.

Preprint. Under review.

We present *content-aware* student models tested in a user study alongside teaching policies for long-term flashcard-based learning. Using pretrained text representations from neural language models [Devlin et al., 2019], our student model infers student familiarity with items without exhaustive testing. Combined with features extracted from practice history similar to Settles and Meeder [2016], our student model has higher accuracy predicting whether a student can correctly answer a trivia question compared to the popular Leitner and SM-2 scheduling methods [Leitner, 1972, Woźniak and Gorzelańczyk, 1994]. To evaluate our methods, we build a web-based flashcard learning app, KAR^{3}L—**K**nowledge **A**cquisition with **R**epresentations and **R**epetition for **R**etention and **L**earning. We pre-populate the app with flashcards from trivia questions and invite members of trivia communities to practice with our app.[1] A four-month user study shows that our content-aware model better captures the dynamics of human learning.

2 Psychology of Memory and Learning

Before discussing our method, we first review relevant concepts from the psychology of learning: mechanisms behind human memory, forgetting curves, and effective long-term study strategies.

2.1 Memory Strength and Forgetting Curves

Forgetting curves describe how memory evolves over time [Ebbinghaus, 1913]. Student models estimate individual forgetting curves of student–item pairs to predict future memory [Mozer et al., 2019]. These estimates of recall probability p over time can be made using a generalized forgetting curve, such as the exponential decay function: $p = be^{-\Delta m}$ [Rubin and Wenzel, 1996], where Δ is the time since last review (i.e., lag time), and m and b are measures of *memory strength*– how factors other than time affect student memory of an item. These factors include (a) student knowledge and ability [Jenkins, 1979], (b) item content and relation to other items (i.e., semantic relationships) [Dunlosky et al., 2013], (c) study type and history [Underwood, 1957], and (d) context of study (e.g., time of day, amount of sleep) [Ebbinghaus, 1913, Jenkins and Dallenbach, 1924]. Student models using a generalized function model m and b to predict an item's individual forgetting curve [Settles and Meeder, 2016, Tabibian et al., 2019].

Identifying the best function for a generalized forgetting curve remains an active research topic [Wixted and Carpenter, 2007, Averell and Heathcote, 2011], but some now question the utility of any general law of forgetting [Roediger, 2008, Erdelyi, 2010]. Proposed functions can account for 98% of variance under experimental settings [Roediger, 2008], but this relies on learning nonsense words tested only once to minimize interference [Ebbinghaus, 1913, Rubin and Wenzel, 1996]. This does not reflect the real world, where students often recall meaningful information multiple times. Equations also assume the moment after study represents the period of highest recall [Roediger, 2008], but long-term encoding processes like "system" *consolidation* can even improve memory over time [Dudai, 2004, McGaugh, 2000]. Under realistic circumstances, adjusting measures of memory strength in generalized forgetting curves is insufficient for student models to accurately estimate a student–item pair's individual forgetting curve [Roediger, 2008]. However, the next section shows how effective study strategies can rely on these idiosyncracies of memory.

2.2 Study Strategies for Optimal Practice

In a review of ten popular learning techniques, Dunlosky et al. [2013] rate *spaced repetition* and *practice testing* as the two best strategies across students and contexts. Spaced repetition spreads out study over increasing intervals as memory strength improves [Kang, 2016]. Practice testing quizzes students on items in low-stakes settings (as opposed to high-stakes one-off summative assessments) [Dunlosky et al., 2013]. Test-taking is particularly effective when feedback is provided [Roediger and Karpicke, 2006], and teachers can use spaced repetition with practice testing strategies such as flashcards Leitner [1972], Kornell [2009].

Semantic relationships affect study strategies. *Interleaved practice*, the study of content from different categories together, is more effective when categories are similar, while its counterpart *blocked*

[1]The word "trivia" is often frowned upon as inaccurate among community members because it suggests memory of arbitrary, useless facts. As reflected in our dataset, most questions are related to academic subjects like science, history, and science. We use this term for lack of a better descriptor.

#	Description
1	Number of successful recalls.
2	Number of failed recalls.
3	Total number of reviews.
4	Average recall rate.

Table 1: Features used in our student model in addition to BERT representations. We construct these features separately for the student, the item, and the combination of the two; in total we have twelve features.

practice is more effective for low-similarity categories [Carvalho and Goldstone, 2014]. Previously studied information affects learning of new information through proactive interference [Underwood, 1957, Roediger, 2008]. Practice testing can improve retention of nontested items [Chan et al., 2006, Chan, 2009], but sometimes it also worsens retention of related items [Anderson et al., 2000, Roediger, 2008]. Existing systems often rely on manual grouping of learning material (e.g., Duolingo lessons) to account for these factors, but without this curation, automated instruction methods must model these semantic relationships to exploit strategies like blocked practice by consecutively showing flashcards about *animal phyla*.

3 Human Learning with ML

This section first reviews machine learning instantiations of student models and teaching policies in automated instruction and then introduces new content-aware approaches for both these tasks.

3.1 Automated Student Models

Settles and Meeder [2016] and subsequent research [Tabibian et al., 2019, Upadhyay et al., 2020] estimate recall probability by training models to determine the measures of memory strength in a generalized forgetting curve. A modified exponential decay curve (Section 2.1) with only one measure of memory strength is a popular base function:

$$f(\Delta, \cdot; \theta) = 2^{-\Delta / h(\cdot;\theta)}, \tag{1}$$

where $\Delta(x_i)$ is the lag time since review, *half-life* function $h(\cdot; \theta)$ represents memory strength [Settles and Meeder, 2016], $\cdot$ is the feature vector for a given input student–item–time triple, and θ are the weights for those features. After practice, these student models adjust half-life to correct recall probability prediction error and to capture change in memory strength from that intervention. This *half-life regression* (HLR) model estimates half-life $\hat{h}_\theta$ by training parameters for manually-crafted features of $\cdot$ to minimize loss in recall probability and half-life.[2] This feature vector allows the model to learn how features of a student–item–pair, such as an item's grammatical features and a student's overall performance, affect memory loss over time.

HLR outperforms a logistic regression model not based on a generalized forgetting curve trained on the same manually-crafted features [Settles and Meeder, 2016]. But as noted in Section 2.1, this dependence on a generalized forgetting curve limits how well it can predict individual forgetting curves. The manually-crafted features used in these student models are limited in capturing semantic relationships. Unlike the student model introduced in the next section, they do not accurately capture the effect of the study strategies dependent on these semantic relationships reviewed in 2.2.

3.1.1 Our Content-Aware Neural Student Model

We introduce a neural network student model that is not forced to follow a general forgetting curve from psychology. This allows our model greater freedom to approximate individual student–item forgetting curves [Cybenko, 1989].

Recall that our goal is to not just learn about students but also needs to represent items. Because we are working in the text domain, we turn to neural language models to create a vector-based

[2]Half-life can use lag time as a feature. But to keep recall estimation f an exponential function of lag-time Δ, half-life h must be either a linear function of or independent of Δ.

representation for each item. These representations are useful for clustering for humans [Bianchi et al., 2020] and computer question answering [Alberti et al., 2019], suggesting they can help capture what humans will know and thus guide what students should study. Thus, we use BERT Devlin et al. [2019] to represent the semantic content [Bengio et al., 2013] of items for review. Compared to existing work using statistical NLP features such as lexeme tags Settles and Meeder [2016], these neural language models provide a much richer representation for the items. Our student model does still maintain similar manually-crafted features (Table 1) to increase interpretability and make comparisons with previous models in ablation studies (Section 4.4). Through these representations, our student model can predict the recall probability for items the student has not seen before; e.g, a student who does not know about the *Iroquois* likely also does not know of the *Seneca*. We can thus better calibrate the difficulty of new items for each student. Existing methods often introduce new items by random sampling because their models can predict little about recall probability of unseen items [Elmes, 2021], but KAR^{3}L's student model predictions mediated by teaching policies can make more informed decisions.

We train this student model using logistic regression with the objective in Equation 2 on previous review history data collected from our platform. Although the groundtruth recall probability p is an unobserved variable, we can get samples from the Bernoulli distribution with parameter p through testing. Given an item, a student, and a point in time, denoted as the triple $\langle x_i, s_j, t \rangle$, the model f with parameters θ outputs a distribution $\hat{p}$ over the two possible outcomes. The objective is to maximize likelihood of the observed outcome y:

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}\left[\hat{p}(y \mid \langle x_i, s_j, t \rangle; \theta)\right] \tag{2}$$

To learn the parameters θ, we can use existing review history data consisting of student–item–time triples Settles and Meeder [2016] and corresponding binary outcomes [Lindsey et al., 2014].

3.2 Automated Teaching Policies

Automated teaching policies can be generally classified as either *interval-based* or *item-based*.

Interval-based teaching policies directly predict the ideal lag time Δ to maximize long-term retention within a budget. After an item review, the student model updates its memory strength estimation and the interval-based policy calculates the next date to review the item based on an updated forgetting curve prediction [Hunziker et al., 2019, Tabibian et al., 2019, Elmes, 2021]. For effective learning, interval-based policies require students to adhere to the recommended schedule.

Item-based teaching policies do not require the student to follow any schedule; instead, it reactively recommends an item based on *when* the student chooses to review [Upadhyay et al., 2020]. When requested at time t, the student model calculates predicted recall probability for all flashcards, both old and new, at time t. Given this data, an item-based policy uses a ranking metric to select the next flashcard for study.

One ranking metric involves selecting the flashcard closest to a perceived level of difficulty, for which recall probability is assumed as a proxy. As preferred difficulty level can depend on student preferences, there is no optimal *target recall probability* value. Existing work have used both 0.85 [Wilson et al., 2019] and 0.5 as targets [Settles and Meeder, 2016].

3.2.1 Our Content-Aware Item-Based Policy

A student model that accurately predicts how all potential interventions affects future memory could determine the most impactful item to study in a situation. However, our item-based policy only uses the student model's prediction of recall probability at the current moment in time. Teaching policies can augment this limited information by incorporating additional heuristics to promote strategies like blocked practice (see Section 2.2). We supplement this threshold system of scoring each flashcard based on distance to the target recall probability predicted by the student model with distance between BERT embeddings to encourage topical coherence during reviews. Concretely, we calculate a weighted average of BERT embeddings of the most recent flashcards (most recent ones are weighted higher), and compute its cosine similarity with all other flashcards. To combine this distance score and the previously discussed recall score, we take a weighted average of the two.

In our experiments, we set the mixing parameter to be 0.5, thus assigning equal importance to the two scores. In principle, this can be customized for each student individually. More importantly,

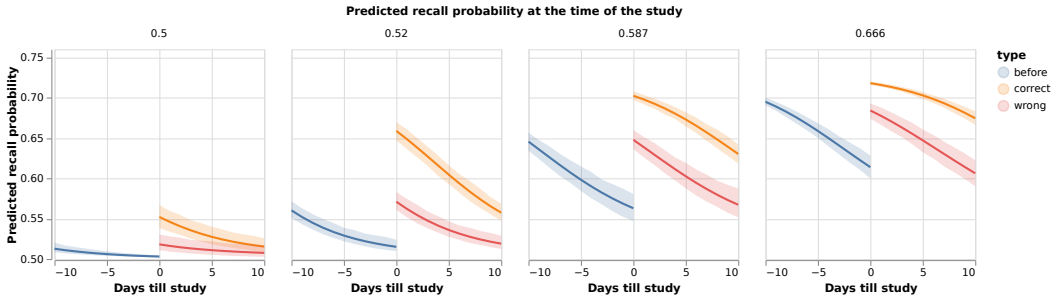

Figure 1: The average forgetting curve ten days before and after a study (which happens at day zero) for both when the user succeed and failed at the recall. Unlike exponential forgetting models, the convexity of our forgetting curve depends on both the current predicted recall and the outcome of the most recent study.

incorporating content representation can enable teaching policies to directly use content-dependent learning strategies (Section 2.2).

4 Experiments

To evaluate student models and teaching policies, we build a publically available web-based flashcard app. This experiment has two goals: to collect a dataset for training and comparing student model designs and to validate our method in a real-world scenario. Unlike previous work where users learn a new language [Settles and Meeder, 2016] or traffic laws [Upadhyay et al., 2020], our flashcard app focuses on the semantically broader domain of trivia. As no flashcard dataset exists in this domain, we repurpose a dataset for online trivia practice [Rodriguez et al., 2019] to cold-start the student model deployed in data collection.

4.1 Why Trivia?

Automated instruction software often target vocabulary acquisition [Kornell, 2009, Settles and Meeder, 2016], where existing spaced repetition flashcard systems are very effective [Altiner, 2011, Bower and Rutson-Griffiths, 2016]. We see evaluating performance for memorizing trivia as a logical next step: It is a popular use case [Gunter], and there are more semantic relationships between study items in subjects ranging from science to literature.

Additionally, trivia learning is an active area of research in natural language processing, with numerous datasets inspired by the trivia community [Boyd-Graber et al., 2012, Joshi et al., 2017, Rodriguez et al., 2019] and models built to solve them [Iyyer et al., 2014, 2015]. The successful application of pretrained representations in trivia question answering [Yamada et al., 2018] suggests it could also encode student knowledge.

4.2 Cold-Starting the Student Model

Although trivia is a suitable testbed for student modeling, no dataset of flashcard learning exists in this domain. We repurpose the Protobowl dataset [Rodriguez et al., 2019] to cold-start our student model. Protobowl is a dataset of people practicing for a trivia game called Quizbowl on `http://protobowl.com`. The dataset contains three millions entries, each including a user ID, a question ID, date and time of the practice, whether the user answered correctly, and some metadata. In Protobowl, the web app randomly chooses from a fixed pool of questions for players to practice, which leads to repeated practices of the same question. The repetition makes the Protobowl dataset a reasonable approximation of a flashcard dataset.

We first filter the dataset by question topic to match the flashcard we host on KAR^{3}L. We only keep good faith entries, and remove entries where the answer is non-sensical. Finally, we filter by a minimum lag time of twelve hours: if the same question is repeated within twelve hours, we keep the last repetition in that period and remove the rest. After these steps we end up with a dataset

Model class	Text repr.	New cards ACC%	AUC	ECE	Old cards ACC%	AUC	ECE
Logistic Regression	BERT	67.6	**.731**	**.119**	**83.1**	**.854**	**.054**
	DistilBERT	**67.8**	.725	.140	81.1	.810	.155
	N/A	62.6	.685	.151	74.5	.729	.162
Exponential Forgetting	BERT	60.2	.624	.142	74.4	.733	.152
	DistilBERT	60.4	.632	.139	73.7	.725	.157
	N/A	60.2	.629	.140	72.0	.729	.159

Table 2: Comparison of student models on predicting recall probability on new and old cards. On three metrics: accuracy (ACC), area under the ROC curve (AUC), and expected calibration error (ECE), our student model combined with BERT representations performs the best.

with roughly 100,000 records. We train our student model and finetune the BERT representations according to Section 3.1.

4.3 Data Collection

We create flashcards by extracting sentences with named entities that occur repeatedly across multi-sentence Quizbowl questions with the same answer in the QANTA question dataset [Rodriguez et al., 2019]. We pair the extracted sentences with the question's associated answer and manually edit these flashcards for clarity. We sort extracted sentence-answer pairs into eleven decks corresponding to subject categories in Quizbowl, which include history, science, and literature. Students using our flashcard app can choose to study these decks and report errors in flashcards.

From August to November 2020, over 400 participants produced over 75,000 study records. Each study record consists of a student ID, flashcard ID, date and time of study, and a binary outcome indicating whether the student successfully recalled the flashcard. We randomly assign 50% of the participants to our method, and divide the rest amongst Leitner and a SM-2 variant;.[3] Participants are unaware of their assignments.

Recruits are drawn from the English-speaking trivia community through advertisements for our study app, and participants were clearly advised before signing up that they were participants of a research study. We further incentivize studying with our app by providing to fifteen participants, totalling $200 in compensation. We collect personally identifiable information in emails, which are not publicized, and usernames only accessible by other app users. In released data, participants are identified exclusively by user id.

4.4 Quantitative Comparison of Student Models

We split the data into training and validation chronologically: *for each user*, the first 75% of the data is used as training, and the last 25% is used for validation. This train/dev split resembles how a student model operates in real life: given existing data on a user, infer future recall probability. To compare student model architectures, we **retrain from scratch** our student model along with baselines. For all models, we use Adam optimizer with learning rate 2e-5 and batch size 64, and train for 10 epochs [Kingma and Ba, 2015].

We perform ablation by removing the BERT representations or replacing them with ones from DistilBERT [Sanh et al., 2019]. DistilBERT uses a set of downstream tasks to finetune and distill BERT, but what's useful for those tasks might not align with what's useful for student modeling. This ablation experiment helps us understand if the student modeling performance is sensitive to the quality of pretrained representations.

Unlike previous work which forces the student model to follow a specific forgetting function (Section 3.1), we consider student modeling simply as logistic regression predicting the binary outcome of each study. Our second ablation experiment compares these two designs. With the same set of

[3]Leitner and SM-2 are popular automated instruction methods reliant on heuristic algorithms for scheduling [Leitner, 1972, Woźniak and Gorzelańczyk, 1994, Elmes, 2021]. To maintain consistency, we reduce student self-evaluation choices in SM-2 from six to two: wrong and right.

features—including the BERT representations—we implement a student model that follows the exponential forgetting curve similar to half-life regression from Settles and Meeder [2016]. Since these models predict recall probability with an exponential function, logistic regression is not applicable. Instead, we follow Settles and Meeder [2016] and train them with mean squared error (MSE).

We use three metrics to compare these models. First, we want models that can reliably predict whether a student can successfully recall a flashcard, so accuracy on this binary classification problem is one metric. But for scheduling flashcards, the prediction confidence of the student model—which we interpret as predicted probability of recall—is also important. So in addition to accuracy, we also consider area under the ROC curve and expected calibration error [Guo et al., 2017].

Table 2 summarizes the comparison of student models. Our logistic regression student model combined with BERT representations performs the best on all three metrics. Replacing BERT with DistilBERT leads to a small decrease in performance, suggesting that student modeling is indeed sensitive to the textual representation, and that the set of downstream tasks use in DistilBERT are not helpful for our student modeling task. As expected, removing textual representations from the model further worsens calibration error. Additionally, our student models trained with logistic regression performs better than the exponential forgetting counterparts. But these metrics cannot tell the full story of how these student models differ [Pelanek, 2015], so next we make some qualitative analyses.

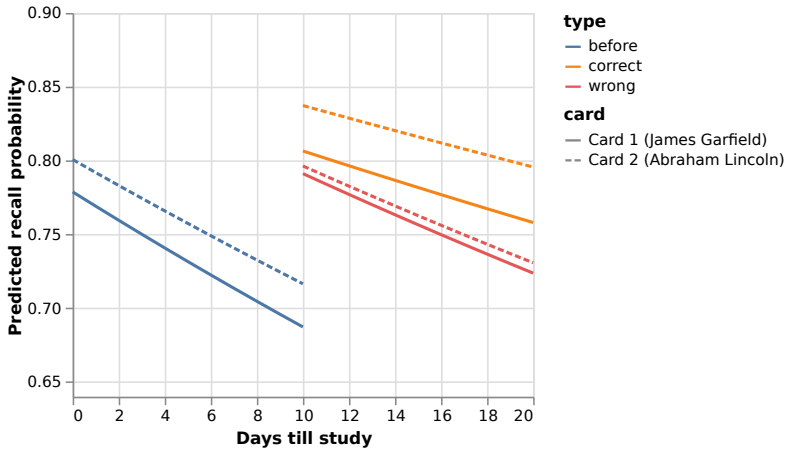

Figure 2: Change in forgetting curve for a semantic related item. Card 1 (James Garfield) is studied at day 0 and again at day 10. When Card is recalled successfully at day 10, the recall probability of the semantically related Card 2 (Abraham Lincoln) increases despite not being studied.

4.5 Visualizing the Forgetting Curves

Unlike existing work, our student model does not follow an explicit forgetting curve. To visualize how our model predicts the decay of memory over time, we repeated query the model for recall probability predictions while changing the date of the query. In Figure 1, we randomly subsample 10% of the validation data, divide them into bins by the predicted recall at the time of the study, and visualizes the forgetting curve ten days before and after. To understand how our model adjusts predictions based on the most recent study, we visualize the forgetting curve for both hypothetical outcomes: when the user succeeds and fails at recall. This figure highlights two important distinctions between our model and existing ones.

Flexible forgetting curve The shape of our forgetting curve—in particular the convexity—depends on both the current predicted recall and the result of most recent study (which happens at day zero in the figure). It is impossible for student models based on proposed functions from psychology like the exponential decay curve (Section 2.1) to adjust in this way: the curve will always be convex. This indicates our student model is more adaptable to variance in individual forgetting curves.

Imperfect memory Existing models assume that memory is perfect immediatedly after a study, so the predicted recall always starts from 100% after a study and decay over time. Our model learned

not to make this assumption: the predicted recall, even immediately after a study, never goes to 100%. Rather, the model adjusts its prediction based on both the predicted recall before study and the result of the study. This behavior reflects how circumstances and study strategies (Section 2.2) can impact levels of memory consolidation (Section 2.1).

Oor model is also able to capture adjust individual forgetting curves of semantically related items. We demonstrate this in Figure 2 with *James Garfield* and *Abraham Lincoln* example from Section 1.

4.6 User Survey Feedback

To gauge user experience and gather feedback, we design a twenty-six question survey with a mixture of open-ended, multiple-choice, and 5-point Likert scale questions. We emailed the survey to registered users and received twenty-six verified responses. Among the respondents, twenty were from the experimental group and used the KAR^{3}L scheduler; the other six were from the control group, three used SM-2 and three used Leitner. Due to limited sample size from the control group, we do not draw any conclusion regarding the comparison between different scheduling methods, and instead focus on suggestions for improving KAR^{3}L. We highlight two takeaways:

- **The desired level of difficulty** varies across users. Some users feel that the cards are too easy; this is potentially also caused by KAR^{3}L reviewing cards too frequently.
- **Topical coherence**. Despite KAR^{3}L using a topic-related term in its teaching policy scoring function, not all users notice whether the presentation of cards resemble blocked practice.

An inability to adjust for user preferences emerges as a main limitation of both KAR^{3}L and existing spaced repetition methods such as Leitner and SM-2. Luckily, as noted in Section 5.3, the formulation of KAR^{3}L can enable customization of difficulty level by the user.

5 Discussion and Future Work

Our experiments show promise for better learning predictions with content-aware student models. This section reviews our collected dataset, evaluation limitations, as well as related and future work.

5.1 A Long-Term Learning Dataset

Future student models can be evaluated on the dataset collected in our user study. Unlike existing datasets [Settles and Meeder, 2016, Bienstman, 2020], we include both item content and repeated studies of a single item over a long-term period. These are important characteristics of flashcard study that facilitate more realistic long-term evaluation [Kornell, 2009].

Our dataset is gathered from the deployed student models and teaching policies in this study. While this affects what items receive more or less attention in the dataset that models are evaluated on, this is unavoidable because there must always be a policy deciding what items are shown. More data from different student model and teaching policy combinations can reduce blindspots in study data.

5.2 Evaluating Teaching Policies and Human Learning

One goal of automated instruction is to improve learning efficiency: to learn more under a limited budget of time and effort. However, learning efficiency must also be evaluated alongside greater goals for learning. Student models may reflect biases from pretrained representations, which means that predictions for certain items may be more accurate than for others. How we construct and evaluate teaching policies can either reflect, mitigate, or amplify inequality in what information is taught.

It is therefore difficult to design metrics evaluating teaching policies: When is a flashcard considered learned? How do we measure student effort and variability? While teaching policy goals affect learning efficiency, whether any policy implementation can meet its goals depends on accurate student model predictions. For this reason, we follow convention by focusing on student model performance [Pelanek, 2015, Settles and Meeder, 2016]. To supplement this evaluation, we gather observational feedback about our teaching policies through the user survey. Although the topic-related term in our teaching policy did not result in discernable topical coherence in our survey, similar strategies of supplementing student model predictions with additional heuristics in teaching policies could mitigate harmful biases in flashcard selection.

5.3 Related and Future Work

Item content is only one under-represented aspect of learning in student models, and additional features could yield further improvement. Jenkins [1979]'s tetrahedral model of memory experiments divides factors affecting memory into four groups: subject (age, ability), materials (pictures, sentences), encoding context (setting, activities), retrieval/test type (cued recall, multiple-choice). KAR^{3}L's neural architecture can extend from manually-crafted features to infer more about these four groups, and incorporating BERT representations allows our student model to infer more about how material factors interact with others to affect memory. Time of study and time spent on a flashcard are example features collected in our dataset that could further improve recall probability predictions.

Additionally, our user survey shows that students want more control over the difficulty of scheduled flashcards. While student-chosen settings may not lead to the most efficient studying, adapting to these preferences could boost time spent studying and therefore total learned information. In KAR^{3}L, the teaching policy's recall probability target parameter explicitly determines difficulty. An option for students to customize target recall probability increases automated instruction software adaptability.

6 Conclusion

This paper shows how machine learning methods can improve personalized education through content-aware student models. Contextual embeddings capture semantic relationships between items useful for modeling learning. We report suggestive evidence in our experiment that the resulting content-aware student model is more accurate and efficient than baseline systems.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA Corpora Generation with Roundtrip Consistency. *arXiv:1906.05416 [cs]*, June 2019.

Cennet Altiner. *Integrating a Computer-Based Flashcard Program into Academic Vocabulary Learning*. Master of Science, Iowa State University, Digital Repository, Ames, 2011.

Michael C. Anderson, Elizabeth L. Bjork, and Robert A. Bjork. Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychon Bull Rev*, 7(3):522–530, September 2000. ISSN 1531-5320. doi: 10.3758/BF03214366.

Lee Averell and Andrew Heathcote. The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, 55(1):25–35, February 2011. ISSN 0022-2496. doi: 10.1016/j.jmp.2010.08.009.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2013.50.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *arXiv:2004.03974 [cs]*, April 2020.

Peter Bienstman. The mnemosyne project. https://mnemosyne-proj.org/, 2020.

Jack Victor Bower and Arthur Rutson-Griffiths. The Relationship between the Use of Spaced Repetition Software with a TOEIC Word List and TOEIC Score Gains. *Computer Assisted Language Learning*, 29(7):1238–1248, 2016. ISSN 0958-8221. doi: 10.1080/09588221.2016.1222444.

Jordan L. Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of Empirical Methods in Natural Language Processing*, 2012.

Paulo F. Carvalho and Robert L. Goldstone. Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Mem Cogn*, 42(3):481–495, April 2014. ISSN 0090-502X, 1532-5946. doi: 10.3758/s13421-013-0371-0.

Jason C. K. Chan, Kathleen B. McDermott, and Henry L. Roediger. Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135(4):553–571, 2006. ISSN 1939-2222(Electronic),0096-3445(Print). doi: 10.1037/0096-3445.135.4.553.

Jason C.K. Chan. When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, 61(2):153–170, August 2009. ISSN 0749596X. doi: 10.1016/j.jml.2009.04.004.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Yadin Dudai. The neurobiology of consolidations, or, how stable is the engram? *Annu Rev Psychol*, 55:51–86, 2004. ISSN 0066-4308. doi: 10.1146/annurev.psych.55.090902.142050.

John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychol Sci Public Interest*, 14(1):4–58, January 2013. ISSN 1529-1006, 1539-6053. doi: 10.1177/1529100612453266.

Hermann Ebbinghaus. *Memory: A Contribution to Experimental Psychology.* Teachers College Press, New York, 1913. doi: 10.1037/10011-000.

Damien Elmes. Anki - powerful, intelligent flashcards. https://apps.ankiweb.net/, 2021.

Matthew Hugh Erdelyi. The ups and downs of memory. *American Psychologist*, 65(7):623–633, 2010. ISSN 1935-990X(Electronic),0003-066X(Print). doi: 10.1037/a0020440.

Emily Gunter. Student Recruitment. http://www.pace-nsc.org/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference of Machine Learning*, 2017.

Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. Teaching multiple concepts to a forgetful learner. In *Proceedings of Advances in Neural Information Processing Systems*, 2019.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Association for Computational Linguistics*, 2015.

J. G. Jenkins and K. M. Dallenbach. Obliviscence During Sleep and Waking. *The American Journal of Psychology*, 35:605–612, 1924. ISSN 1939-8298(Electronic),0002-9556(Print). doi: 10.2307/1414040.

James Jerome Jenkins. *Four Points to Remember: A Tetrahedral Model of Memory Experiments*. 1979. ISBN 978-1-317-74979-0 978-1-317-74980-6 978-1-317-74978-3 978-1-315-79619-2 978-1-315-77503-6.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke S. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*, 2017.

Sean H. K. Kang. Spaced Repetition Promotes Efficient and Effective Learning: Policy Implications for Instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19, March 2016. ISSN 2372-7322, 2372-7330. doi: 10.1177/2372732215624708.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Nate Kornell. Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, 23(9):1297–1317, 2009. ISSN 1099-0720. doi: 10.1002/acp.1537.

Sebastian Leitner. *So lernt man lernen: Angewandte Lernpsychologie ein Weg zum Erfolg*. Herder, 1972. ISBN 978-3-451-16265-7.

Robert V. Lindsey, Jeffery D. Shroyer, Harold Pashler, and Michael C. Mozer. Improving Students' Long-Term Knowledge Retention Through Personalized Review. *Psychol Sci*, 25(3):639–647, March 2014. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797613504302.

James L. McGaugh. Memory–a Century of Consolidation. *Science*, 287(5451):248–251, January 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.287.5451.248.

Michael C Mozer, Melody Wiseheart, and Timothy P Novikoff. Artificial intelligence to support human instruction. *Proceedings of the National Academy of Sciences*, 116(10):3953–3955, 2019.

Radek Pelanek. Metrics for Evaluation of Student Models. June 2015. doi: 10.5281/ZENODO.3554665.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*, 2019.

Henry L. Roediger. Relativity of Remembering: Why the Laws of Memory Vanished. *Annu. Rev. Psychol.*, 59(1):225–254, January 2008. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.psych.57.102904.190139.

Henry L. Roediger and Jeffrey D. Karpicke. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychol Sci*, 17(3):249–255, March 2006. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2006.01693.x.

David C. Rubin and Amy E. Wenzel. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103(4):734–760, 1996. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.103.4.734.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the Association for Computational Linguistics*, 2016.

Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993, 2019.

Benton J. Underwood. Interference and forgetting. *Psychological Review*, 64(1):49–60, 1957. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/h0044616.

Utkarsh Upadhyay, Graham Lancashire, Christoph Moser, and Manuel Gomez-Rodriguez. Large-scale randomized experiment reveals machine learning helps people learn and remember more effectively. *arXiv preprint arXiv:2010.04430*, 2020.

L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, October 1980. ISBN 978-0-674-07668-6.

Robert C Wilson, Amitai Shenhav, Mark Straccia, and Jonathan D Cohen. The eighty five percent rule for optimal learning. *Nature communications*, 2019.

John T. Wixted and Shana K. Carpenter. The Wickelgren Power Law and the Ebbinghaus Savings Function. *Psychol Sci*, 18(2):133–134, February 2007. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.1467-9280.2007.01862.x.

Piotr Woźniak and Edward J Gorzelańczyk. Optimization of repetition spacing in the practice of learning. *Acta Neurobiol Exp (Wars)*, 54(1):59–62, 1994. ISSN 0065-1400.

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. Studio ousia's quiz bowl question answering system. *arXiv preprint arXiv: 1803.08652*, 2018.