Using Semantics of Textbook Highlights to Predict Student Comprehension and Knowledge Retention

David Y.J. Kim[1], Tyler R. Scott[1,2],
Debshila Basu Mallick[3], and Michael C. Mozer[1,2]

[1] University of Colorado, Boulder, CO
[2] Google Research, Brain Team, Mountain View, CA
[3] Rice University, Houston, TX

Abstract. As students read textbooks, they often highlight the material they deem to be most important. We analyze students' highlights to predict their subsequent performance on quiz questions. Past research in this area has encoded highlights in terms of where the highlights appear in the stream of text—a *positional* representation. In this work, we construct a *semantic* representation based on a state-of-the-art deep-learning sentence embedding technique (SBERT) that captures the content-based similarity between quiz questions and highlighted (as well as non-highlighted) sentences in the text. We construct regression models that include latent variables for student skill level and question difficulty and augment the models with highlighting features. We find that highlighting features reliably boost model performance. We conduct experiments that validate models on held-out questions, students, and student-questions and find strong generalization for the latter two but not for held-out questions. Surprisingly, highlighting features improve models for questions at all levels of the Bloom taxonomy, from straightforward recall questions to inferential synthesis/evaluation/creation questions.

Keywords: deep embeddings · natural language processing · student modeling · textbook annotation

1 Introduction

As digital textbooks become increasingly common, researchers have the extraordinary opportunity to observe students as they initially engage with unfamiliar material. Eye gaze has been used as one measure of student behavior [8]. Our interest is in using another source of information that students often provide as they read textbooks: *highlighting* of the material deemed to be most important.

From this manner of student engagement, our goal is to infer students' comprehension and knowledge retention. To the degree that this is possible, early interventions can be designed to steer students toward a deeper understanding of the material.

Our team has engaged in several lines of research on this topic. Winchell et al. [15] conducted a laboratory experiment with three passages from a biology text. Participants were asked to read and highlight the material. Following initial reading, they were given a brief opportunity to review the material along with any highlights they chose to make and were then tested on factual questions that spanned all three sections. Winchell et al. found that the pattern of highlights yield small but reliable improvements in predicting a participant's accuracy of a specific quiz question. Moving to an authentic learning environment, Waters et al. [14] and Kim et al. [5] modeled a data set of highlights obtained from students in actual college-level courses using the OpenStax Tutor platform [12]. Waters et al. found that highlighting the sentence that contains the answer to a question is predictive of performance on that question. Kim et al. extended these results to utilize the entire pattern of highlights in a section of the textbook to predict the overall accuracy of a quiz based on the content of the section.

This past research was limited in two important respects. First, models predicting quiz performance were based on a *positional encoding* of highlights. That is, each section of the text was divided into segments—words, phrases, sentences, or fixed length chunks—and a student's highlighting pattern was represented by a binary vector whose elements indicate whether or not each segment had some highlighting. (Continuous encodings were also explored in which each vector element indicated the proportion of words in that segment that had been highlighted.) Positional encodings contain no explicit information about the *content* of material that has been highlighted; they only allow models to discover regularities such as "if a student highlighted sentence 14 but not sentence 28, their accuracy on question 2 should increase." Such regularities will of course not generalize to other sections of text or to other questions from the same section. A key contribution of the present work is to explore a *semantic encoding* of the highlighted and non-highlighted textbook material. The results presented in this paper show that model accuracy is higher with the semantic encoding than the positional encoding.

The second limitation of past research concerns the nature of information that highlights provide. Models based on only the highlighting pattern may succeed because the highlights provide some general information about how skilled or motivated a particular student is, not because they determine whether students have understood the specific material. To address this possibility, our present work uses a simple latent-variable model, the Rasch model [9], as a baseline. The Rasch model assumes that each student has an *ability* (perhaps better characterized as a skill level) and each question has a *difficulty.* Collaborative filtering methods can be used to infer these latent parameters, from which predictions can be made for new students, new questions, and for known students answering known questions which were not part of the model-training corpus. With the

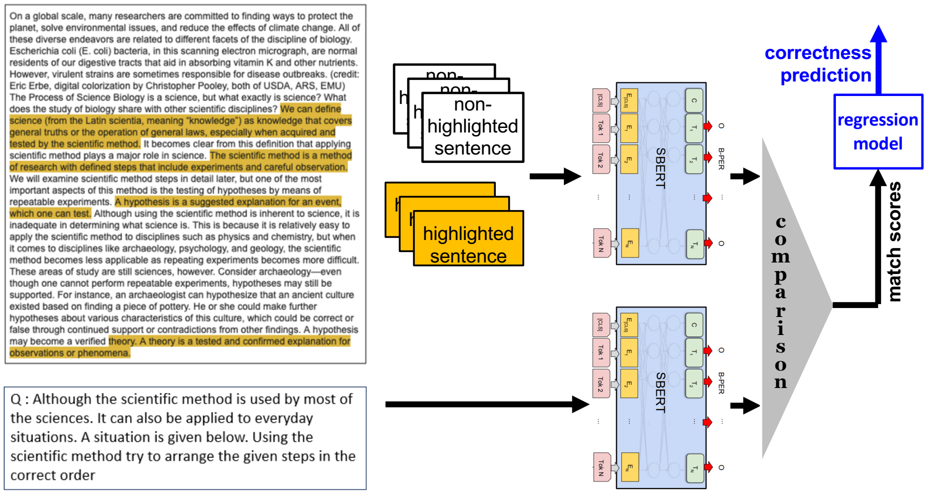
Rasch model as a baseline, we explore whether highlights offer an orthogonal source of information to student ability and question difficulty. We were surprised and pleased to discover that highlights are indeed informative, even when student ability and question difficulty are known.

2 Methodology

2.1 Data

We obtained data from the Openstax Tutor platform [12]. The data were collected from January 1, 2019, through December 31, 2019—spanning two academic semesters—and consist of four different subjects: College Biology, College Physics, Introduction to Sociology, and American History. It is essential to emphasize that these data were collected in a real-world setting, with no control over how the Openstax Tutor platform was administered, and thus, how the data was collected.

The data set consists of 11,134 students, 897 distinct sections, and 830,320 *sessions*, where a session consists of a particular student reading a particular section. We have no further meta-information about the students since the process was completely anonymous, thus we are unable to report or utilize the demographic information about the student sample. For the analysis, we used only



On a global scale, many researchers are committed to finding ways to protect the planet, solve environmental issues, and reduce the effects of climate change. All of these diverse endeavors are related to different facets of the discipline of biology. Escherichia coli (E. coli) bacteria, in this scanning electron micrograph, are normal residents of our digestive tracts that aid in absorbing vitamin K and other nutrients. However, virulent strains are sometimes responsible for disease outbreaks. (credit: Eric Erbe, digital colorization by Christopher Pooley, both of USDA, ARS, EMU) The Process of Science Biology is a science, but what exactly is science? What does the study of biology share with other scientific disciplines? We can define science (from the Latin scientia, meaning "knowledge") as knowledge that covers general truths or the operation of general laws, especially when acquired and tested by the scientific method. It becomes clear from this definition that applying scientific method plays a major role in science. The scientific method is a method of research with defined steps that include experiments and careful observation. We will examine scientific method steps in detail later, but one of the most important aspects of this method is the testing of hypotheses by means of repeatable experiments. A hypothesis is a suggested explanation for an event, which one can test. Although using the scientific method is inherent to science, it is inadequate in determining what science is. This is because it is relatively easy to apply the scientific method to disciplines such as physics and chemistry, but when it comes to disciplines like archaeology, psychology, and geology, the scientific method becomes less applicable as repeating experiments becomes more difficult. These areas of study are still sciences, however. Consider archaeology—even though one cannot perform repeatable experiments, hypotheses may still be supported. For instance, an archaeologist can hypothesize that an ancient culture existed based on finding a piece of pottery. He or she could make further hypotheses about various characteristics of this culture, which could be correct or false through continued support or contradictions from other findings. A hypothesis may become a verified theory. A theory is a tested and confirmed explanation for observations or phenomena.

Q : Although the scientific method is used by most of the sciences. It can also be applied to everyday situations. A situation is given below. Using the scientific method try to arrange the given steps in the correct order

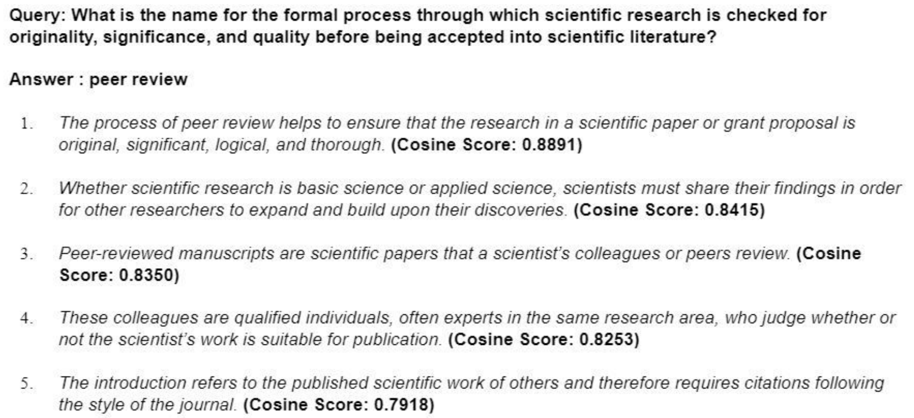
Fig. 1: Sketch of our highlight-based model of student performance. On the left side of the figure is a highlighted passage of text and a specific quiz question. Each of the highlighted and non-highlighted sentences are fed one-at-a-time into SBERT to produce an embedding which is compared with the embedding of the question to determine a match score. The match scores are summarized and fed into a regression model to predict a student's correctness on the given question. Not pictured are latent student-ability and question-difficulty parameters.

the sessions that contain highlights which is 27,019 of the 830,320 sessions. Each section is analyzed independently, and we report mean results across sections. Because the textbooks were electronic, they were revised during the period in which we obtained data. As a result, some sections have multiple versions. We collapsed these revisions together since typically only a few words changed from one version to the next, and it was trivial to align the highlighted fragments.

2.2 Model Design

To capture the semantics of text, we used a pre-trained neural network model: BERT [4]. BERT is a transformer [13] that has produced state-of-the-art results in various natural-language processing tasks. We specifically use Sentence-BERT (SBERT): a modification of BERT that uses a Siamese network structure to derive sentence-level embeddings that can be compared using cosine-similarity [10]. As shown in Figure 1, we predict a student's likelihood of answering a given quiz question correctly by comparing the SBERT embeddings of both highlighted and non-highlighted sentences to the embedding of the question.

In Figure 2, we illustrate the effectiveness of this framework in identifying semantic similarities between sentences from the textbook and quiz questions. The figure shows a sample question from a biology section entitled "The Science of Biology" along with the correct answer to the question. Following the question and answer are the five sentences from the section deemed to be most similar to the question by SBERT. The cosine-similarity score between each sentence and the question is shown in parentheses. In this example, the question is about the definition of peer review. The most related sentence identified by SBERT is a paraphrased definition. The other sentences with high similarity scores are either related to peer review or contain the phrase within the sentence.



Query: What is the name for the formal process through which scientific research is checked for originality, significance, and quality before being accepted into scientific literature?

Answer : peer review

1. *The process of peer review helps to ensure that the research in a scientific paper or grant proposal is original, significant, logical, and thorough.* **(Cosine Score: 0.8891)**
2. *Whether scientific research is basic science or applied science, scientists must share their findings in order for other researchers to expand and build upon their discoveries.* **(Cosine Score: 0.8415)**
3. *Peer-reviewed manuscripts are scientific papers that a scientist's colleagues or peers review.* **(Cosine Score: 0.8350)**
4. *These colleagues are qualified individuals, often experts in the same research area, who judge whether or not the scientist's work is suitable for publication.* **(Cosine Score: 0.8253)**
5. *The introduction refers to the published scientific work of others and therefore requires citations following the style of the journal.* **(Cosine Score: 0.7918)**

Fig. 2: A sample question from a biology section, the correct answer to the question, and the five sentences from the section deemed to be most similar to the question by SBERT.

Representing the semantic similarity between highlights and quiz questions. Here we address several methodological decisions needed to fully specify a predictive model with semantic features. First, we have decided to partition the textbook into sentences [6] and group the sentences in a section into those that have one or more characters highlighted and those that contain no highlights. For each sentence, s, of the section, we obtain an SBERT match score (i.e., cosine similarity) to question q; we denote this match score $B(s, q)$. Since this similarity score would be in the range of $[-1, 1]$, for mathematical convenience and interpretability of model parameters, we rescale this score to the range $[0, 2]$ by adding 1. We thus obtain a set of match scores for highlighted content and a set of match scores for non-highlighted content.

Because the number of sentences—and match scores—in each set varies from student-to-student and section-to-section, we need to recast the two sets of scores into a fixed length vector. A simple approach is to compute the max of the highlighted and non-highlighted sets, resulting in a two-element vector. The maximum score would reflect whether or not the student highlighted the most relevant sentences for a given question. However, the feature is biased in cases where a student highlights excessively. One could instead use the mean score, which would combat over-highlighting, but it's not clear that highlighting material unrelated to the question should make it less likely the student can answer the question. Rather than choosing either the mean or the maximum, we devised a scheme that interpolates between them, and chose a fixed-length vector containing statistics that span the entire range.

If $\boldsymbol{x}$ is a vector of n match scores, and $||\boldsymbol{x}||_p$ denotes the $\mathcal{L}_p$ norm, then $\frac{1}{n}||\boldsymbol{x}||_1$ is the arithmetic mean and $||\boldsymbol{x}||_\infty$ is the maximum. We can define a continuum of norms based on the following relationship:

$$||x||_r \leq n^{\frac{1}{r}-\frac{1}{p}}||x||_p.$$

If we apply this inequality with $p = r + 1$ for all $r = 1, 2, ...$, we obtain the following relation:

$$n^{-1}||x||_1 \leq n^{-\frac{1}{2}}||x||_2 \leq n^{-\frac{1}{3}}||x||_3 \leq ... \leq ||x||_\infty.$$

For a given p, we obtain the following definition of a *highlight match score* or *HMS*:

$$\mathrm{HMS}_{p,q,i} = \left[\frac{1}{n^{\mathrm{H}}} \sum_{s \in S_i^{\mathrm{H}}} B(s,q)^p\right]^{1/p},$$

where S_i^{H} is the set of n^{H} sentences that contain one or more highlights from student i. Because well-matching, non-highlighted sentences might provide additional information, we also construct a score for all the non-highlighted sentences, which we refer to as the *non-highlighted match score* or *NHMS*:

$$\mathrm{NHMS}_{p,q,i} = \left[\frac{1}{n^{\mathrm{NH}}} \sum_{s \in S_i^{\mathrm{NH}}} B(s,q)^p\right]^{1/p},$$

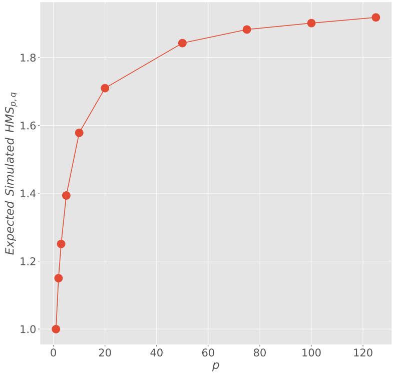


Fig. 3: Expected $\mathrm{HMS}_{p,q,i}$ as a function of p. Vectors of match scores from simulated student i are randomly-sampled with each element, $B(s,q)$, selected from a uniform distribution, $U(0,2)$. Each vector has $n^{\mathrm{H}} = 104$ elements, as that matches the average number of sentences in each section of the text.

where S_i^{NH} is the set of n^{NH} non-highlighted sentences from student i.

As mentioned above, instead of selecting a single value of p to compute HMS and NHMS, we use multiple values. To assist with selecting the values of p, we ran a simulation where we randomly-sampled vectors of match scores, where each match score was selected from a uniform distribution, $U(0,2)$. We then computed the expected HMS for various values of $p \in [1,125]$. The results of the simulation are shown in Figure 3. As expected, $p = 1$ is exactly the mean and $p \to \infty$ approaches the maximum. To approximately span the range, we manually selected $\{1,5,10,100\}$ as the values of p for computing both HMS and NHMS.

Combining the match scores for highlighted and non-highlighted sentences over various values of p, we obtain a parameterized linear model for the overall match:

$$\mathrm{OverallMatch}_{i,q} = \sum_j \alpha_{q,j}\,\mathrm{HMS}_{p_j,q,i} + \sum_j \beta_{q,j}\,\mathrm{NHMS}_{p_j,q,i},$$

where j is an index over a set of norm values $p \in \{1,5,10,100\}$ and $\alpha_{q,j}$ and $\beta_{q,j}$ are free parameters fit to data.

Prediction model. Our prediction model is an extension of the Rasch model [9], a specific instantiation of the classic item-response theory model for students. To formalize the Rasch model, let $y_{i,q} = 1$ if the response from student i to question q is correct. Model predictions are computed as follows:

$$P(y_{i,q} = 1) = \mathrm{logistic}(\theta_i - \gamma_q)$$

where θ_i denotes the latent ability of student i and γ_q denotes the latent difficulty of question q. We refer to the standard Rasch model as A+D since it uses latent parameters for both student ability (A) and question difficulty (D). Our model extends the Rasch model with highlighting features (H), hereafter A+D+H:

$$P(y_{i,q} = 1) = \text{logistic}(\theta_i - \gamma_q + \text{OverallMatch}_{i,q}).$$

For completeness, we compare A+D and A+D+H to four ablated models: (1) A, (2) D, (3) A+H, and (4) D+H.

We perform hierarchical Bayesian inference by placing priors on the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, as well as the parameter matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The priors are as follows:

$$\theta_i \sim N(0, \sigma_\theta^2)\ ,\ \ \gamma_q \sim N(0, \sigma_\gamma^2), \alpha_{q,j} \sim N(0, \sigma_\alpha^2)\ ,\ \ \beta_{q,j} \sim N(0, \sigma_\beta^2)$$

where $\sigma_\theta, \sigma_\gamma, \sigma_\alpha, \sigma_\beta \sim N(0, 2.5)$. All of the models were fit using STAN [2]. We sample four Markov chain Monte Carlo (MCMC) chains each with 4000 samples, and from each chain we remove the first half of samples as burn-in. The remaining samples are then averaged together across the four chains to obtain the estimated parameters, which are then used to compute predictions. We chose hierarchical Bayesian models over a simple maximum likelihood fit to the parameters in order to support principled prediction for new students and to new questions.

We use two performance measures to evaluate models: area under the receiver-operating-characteristic curve (AUC) and the area under the precision-recall curve (PRC). We choose to report PRC in addition to AUC due to an imbalance between correct and incorrect responses to questions in the data. AUC measures a trade-off between sensitivity (or recall) and specificity, neither of which depend on the base rates for each class (i.e., the number of questions correctly answered versus incorrectly answered). PRC, in contrast, computes precision instead of specificity which is sensitive to the base rate of the positive class. In settings where there are many fewer instances of the positive class, PRC assigns more credit to models that successfully classify positive instances (i.e., true positives) [3,11]. We found that our results are consistent with respect to AUC and PRC, but report both for completeness.

3 Results

3.1 Performance within cross-validation settings

We conduct three cross-validation analyses: (1) held-out *student-questions* where the validation set is a random selection of {student, question} pairs, (2) held-out *students* where the validation set contains all questions from a random selection of students, and (3) held-out *questions* where the validation set contains all students from a random selection of questions. In all three cases, we perform five-fold cross validation within each section. The five performance values within each section are averaged, resulting in a single performance metric per section. We then report the mean and standard-error across sections.

Held-out student-questions. In this analysis, the training set typically provides some information about each student and some information about each question. However, it excludes some particular students answering some particular questions. As shown in Figure 4, the three models with highlighting features outperform the corresponding models without highlighting, and the A+D+H model with all features performs the best. Thus, the highlighting features provide distinguishable information from ability and difficulty. We observe that A alone provides the least amount of information, but this is expected since the portion of the training set that constrains each student's ability is far smaller than the portion of the training set that constrains each question's difficulty. Although performance of A+H about matches performance of D, one might suppose that there is redundancy between the two sets of features; however, the superiority of A+D+H over all other models rules out this possibility.

Held-out students. Our next analysis performs cross-validation on students, removing a portion of students from the training set each fold and using them to evaluate the model. This procedure removes any explanatory power of the student ability parameter since at test only the prior distribution is available. As expected (Figure 5), A alone can do no better than chance, yielding an AUC of 0.5, and the models that include ability (purple bars) perform no better than the corresponding models that exclude ability (blue bars). Just as with held-out student-questions, the D+H model outperforms D alone. It is thus reasonable to conclude that the highlighting features provide additional information that can be distinguished from question difficulty.

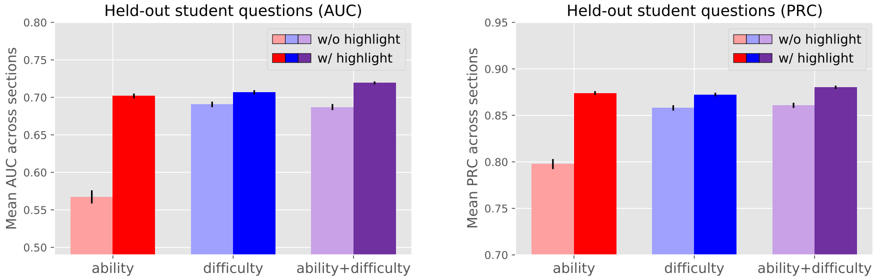


Fig. 4: Results for held-out student-questions with ability, difficulty, and both ability and difficulty features. The darker-colored bars indicate the use of highlighting features in addition to the features listed along the abscissa. Each bar indicates the mean AUC (left) and PRC (right) across sections; error bars reflect ±1 standard-error of the mean, corrected to remove variance due to the random factor [7].

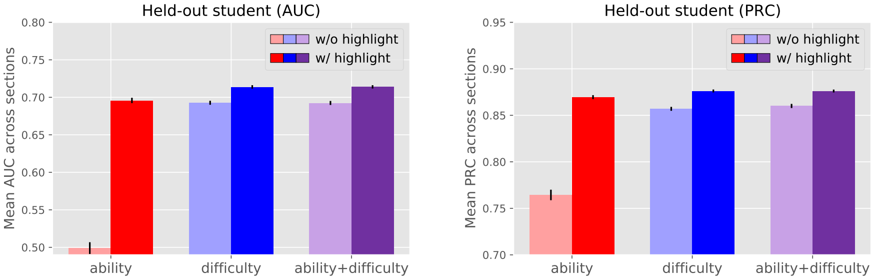


Fig. 5: Results for held-out student models. The plots have identical layout as those in Figure 4. See the caption of Figure 4 for details.

Held-out questions. We performed cross-validation on questions, removing a portion of questions from the training set for each fold and using them to evaluate the model. This procedure removes any explanatory power of the question-difficulty parameter because at test only the prior distribution is available. As expected (Figure 6), D alone can do no better than chance, yielding an AUC around 0.5, and the models that include difficulty (purple bars) perform no better than the corresponding models that exclude difficulty (red bars). The A alone models offers some degree of discrimination; however, none of the models reliably improve when highlighting features are incorporated. This finding is consistent with the laboratory study of Winchell et al. [15] where it was found that with held-out questions, highlighting features did not boost model performance relative to the baseline model (and in fact did somewhat worse due to overfitting). A possible reason for the failure to generalize to new questions is that we train models for each section separately, and each section has relatively few questions. As a result, the model may overfit to the set of questions in the section's training set. We speculate that better generalization to new questions might be obtained

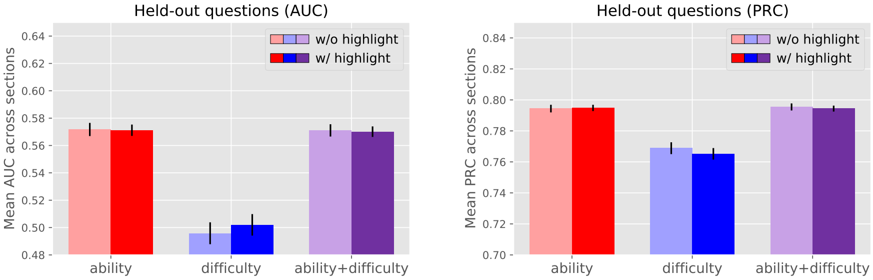


Fig. 6: Results for held-out question models. The plots have identical layout as those in Figure 4. See the caption of Figure 4 for details.

if a single model were trained for all sections rather than using section-specific models. Ongoing simulations are addressing this issue.

While it is disappointing that the current models do not generalize to new questions in a section, this finding does not seriously impact the potential to leverage highlights. When textbooks are designed, the author knows at that point what knowledge should be acquired and correspondingly, what questions should be asked of students. It would be of far greater a concern if models did not generalize to new students; fortunately, our models do this well (Figure 5).

3.2 Performance across levels of conceptual difficulty

In addition to exploring various cross-validation settings, we investigated the performance of both the A+D and A+D+H models across varying levels of conceptual difficulty distinguished by the six levels of the Bloom taxonomy [1]. The taxonomy reflects a continuum from concrete factual questions to abstract reasoning questions; the Bloom levels are: (1) recall, (2) understand, (3) apply, (4) synthesize, (5) evaluate, and (6) create. Waters et al. [14] found that highlights had predictive value only for recall (i.e., Bloom level 1) questions. However, their predictions were based on identifying whether or not a specific critical sentence in the text was highlighted; the information required for questions at higher levels of the Bloom taxonomy are likely to be more diffuse in the text. Thus, the previously used positional encoding of highlights may not have been sufficiently powerful to capture subtle information that the highlights provide.

Because Openstax Tutor had fewer questions at the higher Bloom levels, we clustered Bloom levels. Figure 7 compares A+D models (faint purple) to A+D+H models (dark purple) for three clusters: Bloom level 1, {2,3}, and {4,5,6}. Adding highlighting features improves model performance across all clusters of the Bloom taxonomy. Interestingly, the middle cluster—understand and apply questions—obtains the biggest boost from highlighting features. A

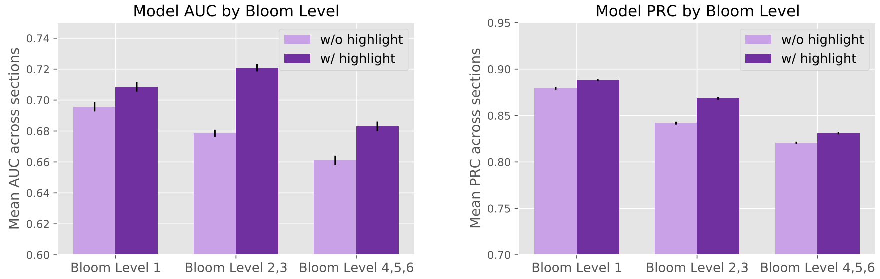


Fig. 7: Held-out student-question results for A+D (lighter-colored bars) and A+D+H (darker-colored bars) across increasing levels of conceptual difficulty, along the abscissa, determined by the Bloom taxonomy. Each bar indicates the mean AUC (left) and PRC (right) across sections; error bars reflect ±1 standard-error of the mean, corrected to remove variance due to the random factor [7].

possible explanation for that is recall questions are so straightforward they do not depend on the complex pattern and semantics of highlights; consequently, the highlighting representation may provide less value. For the third cluster—synthesize, evaluate, and create questions—which require holistic comprehension, our semantic highlighting representation should also be valuable. The predictive power of our models tends to drop for higher levels of the Bloom taxonomy, which we were expecting considering that at higher levels, the complexity of the questions implies that many more factors can come into play in determining student correctness.

3.3 Comparing positional and semantic representations of highlights

In previous work [5,15], we used a positional encoding of highlights. Essentially, we constructed a vector whose elements indicate whether a particular segment of text has been highlighted. We found that providing this high-dimensional vector directly into regression models produced overfitting due to the large number of free parameters. As an alternative, we performed principal-components-analysis (PCA) on the highlighting representation and chose the top k principal components for the highlighting representation. We, in fact, discovered that $k = 1$ worked best generically across sections of text. The previous work is not directly comparable to the present work because it used smaller data sets and Kim et al. [5] evaluated on overall quiz accuracy not individual question accuracy.

We compared the positional highlighting encoding with the encoding developed in this paper and evaluated on individual questions using the current, large data set. As Figure 8 shows, both highlighting representations improve model performance over the baseline A+D model, but augmenting the baseline model with the semantic encoding is superior to augmenting with the positional encoding. We have yet to explore the obvious question of whether augmenting the baseline model with *both* feature sets would further improve model performance.

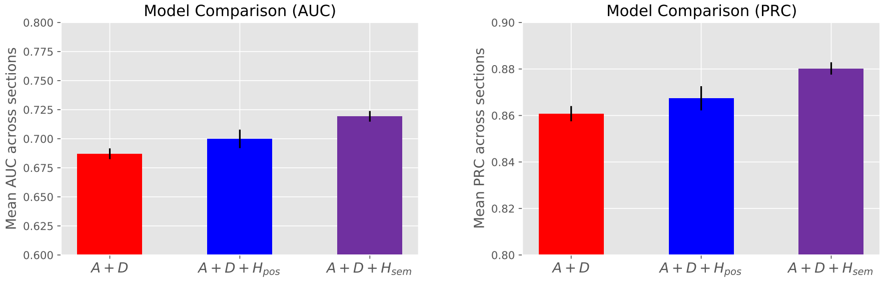


Fig. 8: Comparison of three predictive models with latent ability and difficulty parameters, and optionally using positional or semantic highlighting features, H_{pos} and H_{sem}, respectively. Error bars reflect $+/-1$ SEM, corrected to remove variance due to the random factor [7]

.

4 Conclusions and Future Research

We explored the relationship between student highlighting patterns and question-answering performance using an encoding of highlights based on deep neural network embeddings of text and question content. We found that augmenting a baseline model with this semantic highlighting representation improved predictions of whether a student would answer a specific question correctly. The baseline model is conditioned on latent factors representing student skill level and question difficulty. Our results suggest that highlights provide a source of information that complements these other factors, which may not be surprising in retrospect given that the highlight encoding we used is based on how the particular student interacts with the textbook content that is relevant for the specific question. What is surprising is how effective the SBERT model is in producing embeddings that can be used to judge the similarity of highlighted content to individual questions. We obtained several other key results, including: (1) our models predict well for new students, but not for new questions; (2) our models predict well for all levels of the Bloom taxonomy; and (3) our models that use semantic highlight encodings predict better than models using positional highlight encodings.

From here, there are several potential paths we intend to investigate. First, we should more systematically explore several methodological decisions that we made; in our past work [5], these decisions matter. The assumptions we might question include: whether the correct decomposition of highlights is at the level of complete sentences and not smaller or larger segments; whether a segment of text should be considered highlighted if any portion is highlighted, as opposed to explicitly representing the fraction of the segment highlighted; whether the summary statistics (i.e., values of p) we selected best capture the distribution of highlighted and non-highlighted match scores.

Second, we modeled each section apart from each other section. However, in principle, semantic-highlighting models could apply across multiple sections. Constructing a multi-section model might improve predictions—particularly for held-out questions—because the model would be trained on more data, but it might harm predictions because the weighting of semantic information may vary across sections.

Third, the ultimate goal of our work is not just to predict student performance, but to leverage the predictions to boost student comprehension and retention. Once our investigation of predictive models is complete, the true value of these models to improve student learning can begin.

5 Acknowledgement

This research is supported by NSF awards DRL-1631428 and DRL-1631556. We thank Adam Winchell for helping the initial stage of the research and two anonymous reviewers for their helpful feedback on earlier drafts of this manuscript.

References

1. Bloom, B.S., Krathwohl, D.R., Masia, B.B.: Bloom taxonomy of educational objectives. In: Allyn and Bacon. Pearson Education (1984)
2. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. Grantee Submission **76**(1), 1–32 (2017)
3. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Kim, D.Y., Winchell, A., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G., Mozer, M.C.: Inferring student comprehension from highlighting patterns in digital textbooks: An exploration of an authentic learning platform (2020)
6. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics (2002)
7. Masson, M.E., Loftus, G.R.: Using confidence intervals for graphically based data interpretation. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale **57**(3), 203 (2003)
8. Mills, C., Graesser, A., Risko, E.F., D'Mello, S.K.: Cognitive coupling during reading. Journal of Experimental Psychology: General **146**(6), 872 (2017)
9. Rasch, G.: Probabilistic models for some intelligence and attainment tests. ERIC (1993)
10. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
11. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
12. Stafford, D., Flatley, R.: Openstax. The Charleston Advisor **20**(1), 48–51 (2018)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
14. Waters, A.E., Grimaldi, P.J., Baraniuk, R.G., Mozer, M.C., Pashler, H.: Highlighting associated with improved recall performance in digital learning environment (Submitted)
15. Winchell, A., Lan, A., Mozer, M.: Highlights as an early predictor of student comprehension and interests. Cognitive Science **44**(11), e12901 (2020)