Just Read the Question: Enabling Generalization to New Assessment Items with Text Awareness

Arisha Khan*
Carleton College
khana2@carleton.edu

Nathaniel Li*
Carleton College
lin@carleton.edu

Tori Shen*
Carleton College
khana2@carleton.edu

Anna N. Rafferty
Carleton College
arafferty@carleton.edu

ABSTRACT

Machine learning has been proposed as a way to improve educational assessment by making fine-grained predictions about student performance and learning relationships between items. One challenge with many machine learning approaches is incorporating new items, as these approaches rely heavily on historical data. We develop Text-LENS by extending the LENS partial variational autoencoder for educational assessment to leverage item text embeddings, and explore the impact on predictive performance and generalization to previously unseen items. We examine performance on two datasets: Eedi, a publicly available dataset that includes item content, and LLM-Sim, a novel dataset with test items produced by an LLM. We find that Text-LENS matches LENS' performance on seen items and improves upon it in a variety of conditions involving unseen items; it effectively learns student proficiency from and makes predictions about student performance on new items.

Keywords

educational assessment, text embedding, variational autoencoder

1. INTRODUCTION

Standardized assessments are commonly used in schools, with legal mandates for testing in some countries such as the US [14]. Machine learning approaches to student modeling have been proposed as a way to more efficiently use test data and better align reporting from assessments with the needs of teachers and students (e.g., [3, 18, 20, 21]). These approaches typically require a large amount of data about student performance on each test item. This poses a challenge: the introduction of new test items. Inferences about new items cannot be made until significant student performance data about these items are available. This "cold start" problem, which also exists for typical psychometric approaches, leads to longer assessments in order to field test items [6]. Further, existing models must be retrained to incorporate new items, a resource- and time-intensive process.

To address this challenge, we modify an existing ML model for educational assessment, LENS [3], to leverage text embeddings from a transformer encoder. This model, Text-LENS, can learn more expressive, text-aware item representations.

We also leverage a large language model (LLM) for synthetic test item generation to examine our method's performance in a controlled environment.

We experimentally investigate the effectiveness of Text-LENS and show strong performance when making predictions about performance on seen items based on seen or unseen items. Text-LENS also makes informative predictions about previously unseen items based on the same input information, although performance is not as strong as when the query item is seen. Performance on the simulated dataset demonstrates that Text-LENS incorporates information about both item difficulty and student proficiency, and that it implicitly learns about relationships among items, such as whether they target the same skills. Overall, these results show the potential of this approach as a scalable and effective alternative to assessment models that rely solely on item identity as opposed to content.

2. RELATED WORK

Leveraging text embeddings is increasingly common for educational tasks, including short answer grading [16] and alignment of educational standards [2]. Embeddings of item text have been used to assess properties of assessment items: for instance, to identify similar questions in a large item pool [11] or to classify items based on Bloom's taxonomy [5]. Most closely related to the properties we seek to learn about, several papers have used embeddings to estimate item difficulty (see [1] for a survey). For example, neural networks have been trained on embeddings of the item text and related materials (e.g., a reading passage for a reading comprehension question) [8, 9], and multiple choice item difficulty have been estimated by comparing item and answer choice embeddings [7]. In our work, we explore whether a partial VAE-based architecture can learn about item difficulty and relations among items based on text in order to accurately predict student performance.

Several papers have proposed using question text as part of knowledge tracing. For example, one proposal used LSTMs and attention to learn from both embeddings of item text and historical performance data [17], finding good performance given a sufficiently large training set. Relation-aware knowledge tracing explicitly learns relations among items from embeddings and performance data [15], outperforming [17]. In the context of psychometric approaches

[1]* Equal contribution, listed in alphabetical order.

Arisha Khan, Nathaniel Li, Tori Shen, and Anna Rafferty. Just Read the Question: Enabling Generalization to New Assessment Items with Text Awareness. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 599–603. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870225

like IRT, several recent papers attempt to address the problem of new items through text embedding, either with or without a small amount of preliminary student data on these items [12, 13].

One challenge of exploring these approaches is the lack of publicly available data. Many papers, such as [17], use proprietary data. Pandey et al. [15] scraped several public databases to extract item text; we appreciate that they have made these datasets available, but note quality concerns: many item texts are listed as "TIMEOUT_ISSUE" or the text is insufficient for a human to answer the question (e.g., "Even throwing a dice twice, there are points and 5 5 is the number of chances?"). In assessment-specific work, Benedetto et al. [1] point to item confidentiality and lack of public availability as key barriers to research involving item text. Here, we explore text embeddings for predicting performance on assessments, rather than knowledge tracing tasks, and we use LLM-generated item text to establish proof-of-concept results and to examine the consequences of data structure on model generalization.

3. LEARNING FROM QUESTION TEXT

We introduce Text-LENS, a modification of the LENS model that incorporates question text – instead of learning representations based solely on question identifiers, Text-LENS uses embeddings from a pre-trained transformer encoder.

The base LENS model uses a partial variational auto-encoder (VAE) architecture to produce probabilistic representations of students based on assessment performance [3]. LENS takes two inputs: a set of input items and a set of query items for prediction. Input items include the student correctness on the item. All items are identified to the model by their item id, which is passed through a single feed forward layer to produce an item embedding.

In LENS, the encoder concatenates the student responses and embedded item ids before mapping to the latent space. To make predictions, samples are taken from the latent space, concatenated with the embedded query item ids for prediction, and then decoded to make predictions for query item responses. In Text-LENS, both of the item id embeddings are replaced by text embeddings of the content of these items.

LENS' item id-based embeddings represent relationships between known items but do not encode item content and meaningful interpolations cannot be made between them. In contrast, text-based item embeddings map into continuously meaningful space, even when the particular text of an item has never been seen by the model before. The text-based embeddings also have the potential to encode the difficulty of the items without further training. These qualities improve the generalization of item embeddings and reduce the need for retraining for good performance.

We expect that text embeddings can address several challenges that arise in educational assessment settings. First, text embeddings will allow the model to learn from new items (not present in training) by mapping these new items into the same semantic space. Second, text-awareness will allow the model to align input items with the query item and infer the difficulties of items for better prediction on unseen query items.

4. DATASETS

Testing the effectiveness of Text-LENS requires datasets where item text and student performance data are available. There are limited publicly available datasets that meet these requirements. We thus use both synthetic and real-world data: the synthetic dataset allows for more fine-grained control and checking or assumptions, while the real-world datasets validate these results in a realistic setting.

EEDI: We used data from Task 3 & 4 of the NeurIPS 2020 Education Challenge [19]. It consists of responses to multiple choice math items from two school years on the Eedi platform. Item content is provided as images, which we processed using Optical Character Recognition (OCR). OCR introduced considerable noise: Mathematical formulas are not always accurate, and sequencing of item components was occasionally distorted in the output text.

We filtered the data to include only five distinct skills from the skill taxonomy in order to facilitate experimentation with the relationship between input and query items. After randomly selecting among those skills associated with at least 20 items, the data included $4,641$ students and 214 unique items.

LLM-SIM: This fully synthetic dataset consists of items generated by prompting GPT-4o about specific topics, eliminating the OCR artifacts while focusing on similar topics. The LLM was told to generate ten easy, medium, and hard math test questions for the given subject, and asked to "significantly vary the wording of the questions, such as the type of information included in the question and the phrasing of the task itself." Fourteen topics were randomly selected from level 3 (of 4 total levels) of the Eedi skill taxonomy. For each skill, we prompted the LLM separately about each of its 3-4 subskills. This resulted in 90-120 items per skill.[1] Since later experiments needed only five skills and 40 items per skill, we filtered the data to randomly select only the needed number of items, resulting in a total of 200 items (40 items for each of 5 skills).

After generating item text, we created $50,000$ simulated students and sampled their responses to all items via an IRT model. For each simulated student, we independently sampled one proficiency parameter per skill; proficiency parameters were sampled from $\mathcal{N}(0,1)$, clipped to fall in $[-4,4]$. By structuring the five skill proficiencies to be independent of one another, we can investigate whether Text-LENS correctly learns relationships between items and skills. Items were assigned difficulty parameters of -1.5 (easy), 0 (medium), or 1.5 (hard) based on the LLM's estimated difficulty level when generating the question. For each item, the probability the student would answer correctly was calculated according to a 3-PL IRT model [10] with discrimination parameter 1 and guessing parameter 0.1. To enable a variety of experimental structures, we create simulated responses for all 200 items for every student.

5. EXPERIMENTS

Our experiments aim to assess how effectively Text-LENS leverages item text to improve assessment predictions. We test Text-LENS' ability to extract item information and understand the degree to which it can make generalizations about new (unseen) items. Specifically, we investigated the following research questions:

- **Performance on seen queries:** How well does Text-LENS incorporate information from seen or unseen input to make predictions about performance on seen query items?
- **Performance on unseen queries:** How effectively can Text-LENS make predictions about student performance on unseen query items?

[1] Full dataset available here.

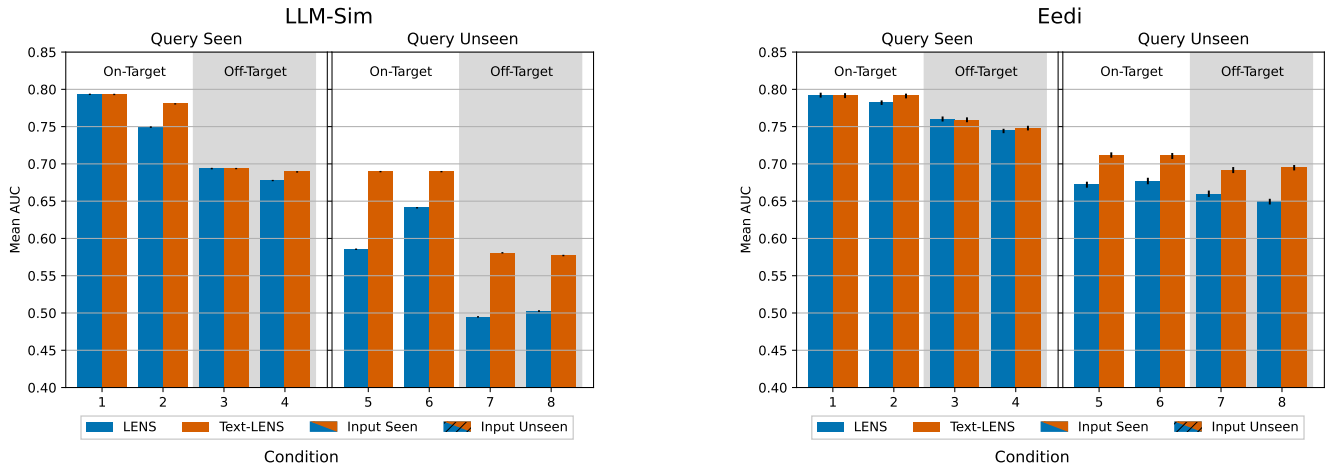

Figure 1: LENS and Text-LENS performance on LLM-Sim (left) and Eedi (right) across all conditions. The error bars represent one standard error over 60 repetitions.

- **Varying relationships between input and query:** How does the relationship between the input items and the query items affect performance predictions?

5.1 Training and Item Embeddings

We evaluate LENS and Text-LENS on Eedi and LLM-Sim. We train on 80% of students and use 10% each for validation and testing. We partitioned the items into halves. Models were trained on one half, reserving the remaining items as unseen.

For these experiments, we used mathBERT as our pretrained transformer encoder to produce the item embeddings for Text-LENS [4]. The original LENS model serves as our performance baseline.

We trained LENS and Text-LENS on each dataset using the model parameters from the original LENS paper, but performed grid search on learning rate (LR), distribution dimensions, encoder hidden dimensions, and accumulator hidden dimensions for each of our datasets and for both LENS and Text-LENS. These parameters were selected for their relevance to the model's ability to represent the number of skills in a dataset. All models were trained using the Adam optimizer for 300 epochs.

Table 1: Parameters used in training of each model.

Dataset	Model	LR	Dist. Dim.	Encoder Hidden Dim.	Accumulator Hidden Dim.
Eedi	LENS	.001	16	30	8
	Text-LENS	.001	16	30	8
LLM-Sim	LENS	.001	64	90	24
	Text-LENS	.005	64	30	24

5.2 Test Setup

We evaluated our models to assess performance with unseen items. For each test student, the model received a query item to predict and a set of input items containing student responses. The skill associated with the query item was designated as the target skill. If the input items were selected to be of the same skill as the query, we consider them to be on-target.

As shown in Table 2, our experimental design varied three factors:

Table 2: Experimental conditions varying which items are previously seen and the relationship between input and query items.

Conditions	Input Items				Query
	Target Skills		Other Skills		
	Seen	Unseen	Seen	Unseen	
1	19	0	0	0	Seen
2	0	19	0	0	
3	0	0	19	0	
4	0	0	0	19	
5	19	0	0	0	Unseen
6	0	19	0	0	
7	0	0	19	0	
8	0	0	0	19	

(1) whether the input items were seen or unseen during training, (2) whether the query item was seen or unseen during training, and (3) whether the input items' skill was on- or off-target. These conditions allows us to assess how effectively text embeddings enable generalization to novel questions compared to ID-based representations across both synthetic and real-world datasets.

6. RESULTS

We find that Text-LENS consistently matches or surpasses LENS, highlighting the potential of text-awareness to enhance test inference flexibility and performance.

6.1 Performance on seen queries

When all input and query items have previously been seen in training, Text-LENS and LENS show very similar performance (Figure 1, conditions 1 and 3). This parity indicates that text-based embeddings can effectively capture at least as much information as ID-based embeddings for items that have been seen during training.

Text-LENS uses the embeddings of unseen input items to effectively make inferences about previously seen query items: it performs as well (Eedi) and nearly as well (LLM-Sim) in condition 2 as in condition 1 (Figure 1). Meanwhile, LENS' performance in condition 2 is somewhat degraded. Both models still extract

meaningful information from unseen input items since the items still generally characterize the student's skill. This insight explains LENS' strong performance even when the input is unseen on condition 2: higher performance on the input is strongly correlated with success on the query item, mirroring the pattern from training.

6.2 Performance on unseen queries

In all conditions with unseen queries, Text-LENS achieves a higher AUC than LENS (right panels in Figure 1). However, inferences about unseen query items are less accurate than inferences about seen query items: for instance, when input items are seen but query unseen (condition 5), Text-LENS' AUC drops 0.08 on Eedi and 0.10 on LLM-Sim compared to when query items were seen (condition 1). LENS falls more sharply, decreasing 0.12 on Eedi and 0.20 on LLM-Sim. Unseen queries are associated with a larger drop in performance than unseen input (condition 1 vs. 5 compared to condition 1 vs. 2). This larger penalty likely reflects the fact that with unseen input, the model receives both an item identifier and student performance on that item, but only the identifier is available for the query.

6.3 Varying input relevancy

In the previous comparisons, the target skill of the input items aligned with that of the query item. The models' inferences depend on both estimates of student proficiency on this skill and estimates of the difficulty of the query item. When the input items are aligned with a different skill than the query item, the model must fall back to using estimates of query item difficulty and any learned relationships among skills. In LLM-Sim, skills are independent, whereas relationships among skills may exist in Eedi.

Off-target query items consistently result in lower performance compared to the analogous on-target query conditions (see Figure 1), especially for LLM-Sim. Text-LENS seems to learn about query item difficulty: for conditions 7 and 8, LENS performs no better than random, but Text-LENS achieves an AUC of about 0.58. Since the input items are off-target and the query item is unseen in training, Text-LENS could only have made its inference by extracting difficulty from text, and its better-than-random performance indicates that it is doing so. To further confirm, we simulated and retrained the models on LLM-Sim where the relationship between item text and difficulty was randomized and observed the Text-LENS' performance drop to chance.

While there is still some performance drop for off-target query items on Eedi, this drop is much more modest. This likely reflects that, while the skill alignment between input and query is relevant to inference, there are also strong dependencies between skills that allow off-target skills to provide good information about the student's response to the query item.

6.4 Performance across different datasets

While the relative performance between Text-LENS and LENS for any particular condition is preserved across both datasets, we find some performance differences between our datasets.

Surprisingly, on LLM-Sim, LENS performs better when making inferences about unseen query items with unseen input items (condition 5) than with seen (condition 4). Here, the seen input items could be misleading, since the item ID-based embedding method can create meaningful embeddings for the input, but any relationship between the input and query embeddings is spurious. In contrast, LENS' input item embeddings are unreliable in condition 5 – just as in condition 2. Here, the model may infer based only on a student's general performance and obtain good results by doing so.

This increase in LENS' performance from condition 4 to 5 is not present on Eedi. Here, due to dependencies between skills, it is not as important for the model to understand the relationships between the input and query. This also explains the smaller performance change between conditions 1 and 2 on Eedi than on LLM-Sim.

Text-LENS performs consistently across both datasets, with no unpredictable behavior like LENS' performance jump from condition 4 to 5, except for the greater impact of off-target items in LLM-Sim due to complete skill independence.

7. DISCUSSION

While traditional assessment models require extensive historical performance data for each new item, our text-aware approach allows immediate integration of unseen questions, improving the "cold start" problem. Text-LENS could significantly reduce the need for field testing new items before operational use, shortening assessment length and decreasing student testing time. Additionally, this approach enables dynamic adaptation of assessments—instructors can modify questions or introduce entirely new ones while maintaining predictive accuracy about student performance. Our experiments demonstrate that text embeddings effectively encode item difficulty and skill relationships directly from question content, allowing for more flexible assessment design and reducing the resource-intensive retraining typically required when item pools evolve.

Natural directions for future work include further investigation into real-world performance and the value of LLM-based simulated data. Our evaluation on LLM-Sim indicates the ability of text embedding to encode difficulty, but it is unclear to what degree real-world questions encode difficulty in text and how that differs from an LLM specifically prompted to produce questions at different difficulties. This evidence is in alignment with some prior work (see Section 2), but that work was not focused on ML models for assessment and was typically focused on non-math subjects. A high-quality assessment dataset with test content and either professionally annotated difficulties or difficulties inferred by an IRT-based system would help to investigate the degree to which Text-LENS can recover item difficulty from text.

Our evaluation of real-world data was limited to Eedi, for which test content only exists as screenshots of items. Thus, our item embeddings were produced from noisy and sometimes incorrect OCR. Furthermore, any graphical or spatial information included in the question was lost. Future work could attempt to incorporate multimodal items; prompting a multimodal LLM to summarize the item in text might be one path forward. A further limitation is our focus solely on math problems. In Eedi, this might lead to high dependencies among skills compared to some other subjects. Math questions may rely more on diagrams and have less rich text, so Text-LENS might exhibit stronger performance on subjects where language features more prominently, such as reading comprehension.

The text of exam questions provides a rich source of information that can be applied toward response inference by leveraging text embeddings. While further evaluation is needed prior to deployment for real-world applications, the strong performance of Text-LENS across a variety of realistic use cases demonstrates the feasibility of using text embeddings for ML-based assessment models.

8. ACKNOWLEDGEMENTS

We thank Aadi Akyianu for their initial work on the project, and Jared Arroyo-Ruiz, Geoffrey Jing, and Nhi Luong for prior involvement in this research project. We also thank Carson Cook and S. Thomas Christie for early discussions about the project idea and direction.

9. REFERENCES

[1] Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., Turrin, R.: A survey on recent approaches to question difficulty estimation from text. ACM Computing Surveys **55**(9), 1–37 (2023)

[2] Butterfuss, R., Doran, H.: An application of text embeddings to support alignment of educational content standards. Educational Measurement: Issues and Practice **44**(1), 73–83 (2025)

[3] Christie, S.T., Johnson, H., Cook, C., Gianopulos, G., Rafferty, A.N.: Lens: Predictive diagnostics for flexible and efficient assessments. In: Proceedings of the Tenth ACM Conference on Learning @ Scale. p. 14–24. L@S '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3573051.3593392

[4] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)

[5] Gani, M.O., Ayyasamy, R.K., Sangodiah, A., Fui, Y.T.: Bloom's taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. Education and Information Technologies **28**(12), 15893–15914 (2023)

[6] Haladyna, T.M., Rodriguez, M.C.: Developing and validating test items. Routledge (2013)

[7] Hsu, F.Y., Lee, H.M., Chang, T.H., Sung, Y.T.: Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. Information Processing & Management **54**(6), 969–984 (2018)

[8] Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., Hu, G.: Question difficulty prediction for READING problems in standard tests. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. p. 1352–1359. AAAI'17, AAAI Press (2017)

[9] Lin, L.H., Chang, T.H., Hsu, F.Y.: Automated prediction of item difficulty in reading comprehension using long short-term memory. In: 2019 International Conference on Asian Language Processing (IALP). pp. 132–135. IEEE (2019)

[10] Van der Linden, W.J., van der Linden, W.: Handbook of item response theory, vol. 1. CRC Press New York (2016)

[11] Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y., Hu, G.: Finding similar exercises in online education systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1821–1830 (2018)

[12] Ma, H., Xia, A., Wang, C., Wang, H., Zhang, X.: Diffusion-inspired cold start with sufficient prior in computerized adaptive testing. arXiv preprint arXiv:2411.12182 (2024)

[13] McCarthy, A.D., Yancey, K.P., LaFlair, G.T., Egbert, J., Liao, M., Settles, B.: Jump-starting item parameters for adaptive language tests. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 883–899. Association for Computational Linguistics (Nov 2021)

[14] National Education Association: ESSA and Testing (2020), https://www.nea.org/resource-library/essa-and-testing, accessed: 2025-03-13

[15] Pandey, S., Srivastava, J.: Rkt: relation-aware self-attention for knowledge tracing. In: Proceedings of the 29th ACM international conference on information & knowledge management. pp. 1205–1214 (2020)

[16] Putnikovic, M., Jovanovic, J.: Embeddings for automatic short answer grading: A scoping review. IEEE Transactions on Learning Technologies **16**(2), 219–231 (2023)

[17] Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., Hu, G.: Exercise-enhanced sequential modeling for student performance prediction. vol. 32 (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11864

[18] Wang, Z., Gu, Y., Lan, A., Baraniuk, R.: VarFA: A variational factor analysis framework for efficient bayesian learning analytics. In: Proceedings of the 13th International Conference on Educational Data Mining. pp. 696–699 (2020)

[19] Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J.M., Turner, R.E., Baraniuk, R.G., Barton, C., Jones, S.P., Woodhead, S., Zhang, C.: Diagnostic questions: The NeurIPS 2020 education challenge. arXiv preprint arXiv:2007.12061 (2020)

[20] Wu, M., Davis, R.L., Domingue, B.W., Piech, C., Goodman, N.: Variational item response theory: Fast, accurate, and expressive. In: Proceedings of the 13th International Conference on Educational Data Mining. pp. 257–268 (2020)

[21] Zhai, X., Yin, Y., Pellegrino, J.W., Haudek, K.C., Shi, L.: Applying machine learning in science assessment: a systematic review. Studies in Science Education **56**(1), 111–151 (2020)