Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation

Keith Cochran, Clayton Cohn, Jean-François Rouet, Peter Hastings

Improving Automated Evaluation of Student Text Responses using GPT-3.5 for Text Data Augmentation

Keith Cochran[1, [0000−0003−0227−7903]], Clayton Cohn[2, [0000−0003−0856−9587]], Jean Francois Rouet[3, [0000−0002−0368−7691]], and Peter Hastings[1, [0000−0002−0183−001X]]

[1] DePaul University, Chicago IL 60604, USA
[2] Vanderbilt University, Nashville TN 37240, USA,
[3] Université de Poitiers, 86073 Poitiers Cedex 9, France
kcochr11@depaul.edu

Abstract. In education, intelligent learning environments allow students to choose how to tackle open-ended tasks while monitoring performance and behavior, allowing for the creation of adaptive support to help students overcome challenges. Timely feedback is critical to aid students' progression toward learning and improved problem-solving. Feedback on text-based student responses can be delayed when teachers are overloaded with work. Automated evaluation can provide quick student feedback while easing the manual evaluation burden for teachers in areas with a high teacher-to-student ratio. Current methods of evaluating student essay responses to questions have included transformer-based natural language processing models with varying degrees of success. One main challenge in training these models is the scarcity of data for student-generated data. Larger volumes of training data are needed to create models that perform at a sufficient level of accuracy. Some studies have vast data, but large quantities are difficult to obtain when educational studies involve student-generated text. To overcome this data scarcity issue, text augmentation techniques have been employed to balance and expand the data set so that models can be trained with higher accuracy, leading to more reliable evaluation and categorization of student answers to aid teachers in the student's learning progression. This paper examines the text-generating AI model, GPT-3.5, to determine if prompt-based text-generation methods are viable for generating additional text to supplement small sets of student responses for machine learning model training. We augmented student responses across two domains using GPT-3.5 completions and used that data to train a multilingual BERT model. Our results show that text generation can improve model performance on small data sets over simple self-augmentation.

Keywords: data augmentation · text generation · BERT · GPT-3.5 · educational texts · natural language processing

1 Introduction

Researchers in educational contexts investigate how students reason and learn to discover new ways to evaluate their performance and provide feedback that promotes growth. Intelligent learning environments (ILEs) for K-12 students are designed to incorporate inquiry-based, problem-solving, game-based, and open-ended learning approaches [17,21,23]. By allowing students to choose how they approach and tackle open-ended tasks [37], they can utilize the resources available in the environment to gather information, understand the problem, and apply their knowledge to solve problems and achieve their learning objectives. At the same time, ILEs monitor students' performance and behavior, allowing for the creation of adaptive support to help students overcome challenges and become more effective learners [7,2,33].

Some research in this field aims to understand the factors that impact learning in various contexts. One area of study is centered on national and international literacy standards [1], which mandate that students should be able to think critically about science-related texts, understand scientific arguments, evaluate them, and produce well-written summaries. This is crucial for addressing societal issues such as bias, "fake news," and civic responsibility. However, achieving deep comprehension of explanations and arguments can be difficult for teenage students [25]. Additionally, research in discourse psychology suggests that students' reading strategies are shaped by their assigned reading task and other contextual dimensions [8]. For example, prior research has shown that students generate different types of inferences when reading as if to prepare for an exam compared to reading for leisure [9]. Similarly, students' writing is influenced by their perception of the audience [12].

Student responses in educational settings usually have a specific structure or purpose, which aligns with the grading criteria and demonstrates the student's level of understanding of the material. Natural Language Processing (NLP) techniques like sentence classification can be used to analyze student performance and provide feedback quickly [19]. *BERT-based* models have revolutionized the NLP field by being pre-trained on large datasets such as Wikipedia and BooksCorpus [15], giving them a deep understanding of language and how words are used *in context*. These models can then be fine-tuned for specific tasks by adding an output layer and training it with a smaller labeled dataset.

One common approach to improve models' performance with limited data is data augmentation [30]. This technique is commonly used in other fields of AI, such as computer vision. Attempts have been made to apply data augmentation techniques to text data [11], but it is more challenging because small changes in the text can produce bigger changes in the meaning, leading to errors in model training. Some current data augmentation techniques for text data involve modifying original responses, such as misspelling words or replacing them with similar words [34].

In this paper, we investigate text generation using three different "temperatures" and compare the results to a baseline measurement and a self-augmentation method, where the original data set is replicated to increase the training data.

This technique of self-augmentation has been successful in previous research [13], and similar methods have been applied to computer vision with improved model performance [29]. We aim to determine the appropriate level of augmentation and establish a baseline measurement for comparison when additional augmentation techniques are applied.

2 Background and Research Questions

Data sets in educational contexts can sometimes be large, but when they are comprised of students' hand-generated responses, they tend to be on the order of at most few hundred responses. The amount of data obtained is sometimes a function of the nature of the tests. Modern machine learning models come pre-trained on various data sets. However, in order to improve performance on a given downstream task, these models need to be fine-tuned using labeled data [36]. Although some of these models can be good at zero-shot or few-shot learning [35], they are designed to allow further fine-tuning to improve performance for specific tasks when sufficient training data in both quantity and quality is available [18].

These educational data sets are also often imbalanced, meaning each label does not have equal representation. Machine learning models perform better when the data is close to being balanced across labels [28]. Data augmentation has improved model performance in image processing [30]. However, that process does not translate directly to text-based models. Simple replication of the data can be used and is referred to as self-augmentation. Looking at techniques beyond self-augmentation, [6] describes a taxonomy and grouping for data augmentation types. Cochran et al. showed that augmentation using masking, noise, and synonyms can improve classification performance [14]. This study continues that research by exploring augmentation using a generative AI method.

Recent studies have used text generation to improve classifier performance by augmenting data to create additional training data artificially [31,27]. The intent is to address the imbalance in data sets and allow smaller data sets to acquire larger data volumes to aid model training. Several survey papers on text augmentation break down the various types of data augmentation currently being researched [5,22,16]. In the generative method of text augmentation, artificial student responses are generated using a predictive model that predicts the response given a text prompt as input.

The OpenAI API performs NLP tasks such as classification or natural language generation given a prompt. OpenAI provides an interface for the Generative Pretrained Transformer 3.5 (GPT-3.5) [10], one of the most powerful language models available today [26]. A recent study has shown improvement for short text classification with augmented data from GPT-3.5, stating that it can be used without additional fine-tuning to improve classification performance [3]. Additionally, [6] note that GPT-3.5 is the leading augmentation method among recent papers and may even be able to replicate some instances whose labels are left out of the data set (zero-shot learning).

The student response data sets contain labels for each response corresponding to a hand-graded value on a grading rubric. Transformer-based NLP models, such as BERT [15] and GPT-3.5 [10], are now the industry standard for modeling many NLP tasks. Previous research by [14] shows that BERT-based transformers work well for text classification of student responses to STEM questions. Therefore, we are using GPT-3.5 for augmentation and continue to use BERT-based models for classification. Since we have two data sets in two languages, English and French, we use a BERT-based multilingual model as the classifier of choice.

RQ 1: Can artificially-generated responses improve base model classification performance? Our hypothesis **H1** is that additional augmented data will improve model performance for smaller data sets. Determining how large a data set needs to be before it would not require data augmentation is out of the scope of this study. Here, we are determining if augmentation will work for these data sets at all (i.e. can we reject the null hypothesis that the models perform the same with and without augmented data).

RQ 2: Can artificially-generated responses outperform self-augmentation when used for training models for sentence classification? Our hypothesis **H2** is that artificially-generated responses will outperform the self-augmentation method because they are not simple copies of the data, so more of the domain is likely to be filled with unique examples when creating the augmented data space.

RQ 3: Does temperature sampling of the artificially-generated student responses affect model performance? Recall the temperature variable for the OpenAI API allows for altering the probability distribution for a given pool of most likely completions. A lower value creates responses almost identical to the prompt text. A higher value (up to a maximum of 1) allows the model to choose more "risky" choices from a wider statistical field. **H3** proposes that augmenting the data with slightly more risky answers, equating to a temperature of 0.5, will provide the best performance in general. Until we test, we do not want to speculate if a number closer to 1.0 would improve performance or not. We hypothesize the temperature setting of 0.5 would be on the low side.

RQ 4: Does performance ultimately degrade when the model reaches a sufficient level of augmentation? It can be assumed that any augmentation would encounter overfitting, where model performance begins to degrade at some point [13]. **H4** is that the performance will degrade with additional augmentation after a peak is reached. **H5** is that the performance will degrade more slowly with higher temperature augmented data sets and thus support the idea that more risk involved in generated responses is better for larger amounts of augmentation.

3 Methods

3.1 Data Sets

Two data sets were obtained for this study. The first data set is from a discourse psychology experiment at a French university where 163 students were given an article describing links between personal aggression and playing violent video

games. The participants were asked to read the article and write a passage to either a friend in a "personal" condition or a colleague in an "academic" condition. Our evaluation was around whether or not they asserted an opinion on the link between violent video games and personal aggression. The label quantities from the data set are shown in Table 1. The majority label quantity, "No Opinion", is shown in bold. The rightmost column gives the Entropy measure for the data, normalized for the four possible outcomes. A dataset which is balanced across labels would have an entropy value near 1.

Table 1. French Student Response Data Split for the Opinion Concept

Label	No Opinion	Link Exists	No Link	Partial Link	Normalized Entropy
Count	**118**	7	13	25	0.619

The second data set was obtained from a study [20,4,24,37] on students learning about rainwater runoff with responses from 95 6^{th}-grade students in the southeastern United States. Responses were given in the English language.

Each of the six concepts was modeled individually as a binary classification task. Student responses that included the corresponding concept were coded as **Present**. Responses were otherwise coded as **Absent**. As previously mentioned, many small educational data sets are imbalanced. Table 2 shows the label quantities indicating the scarcity of data and the degree of imbalance in the dataset.

Table 2. Rainwater Runoff Student Response Data Split per Question

Concept	Absent	Present	Entropy
1	10	**85**	0.485
2a	25	**70**	0.831
2b	**64**	31	0.911
3a	44	**51**	0.996
3b	**73**	22	0.895
3c	**57**	38	0.971

3.2 Augmentation Approach

The label quantities shown in Tables 1 and 2, along with the normalized entropy, have the **majority quantity of reference** for that particular data set shown in bold. A balanced data set would have equal quantities across all labels and normalized entropy values at or near 1.0. We define an augmentation level of 1x when all labels have the same quantity as the majority quantity of reference for

that data set. All data sets after that was augmented in multiples of the majority quantity up to 100x, or 100 times the majority quantity of reference for that data set. We generated data using GPT-3.5 (model "text-curie-001") with the prompt "paraphrase this sentence" and inserted an actual student response to fill in the rest of the language prompt. The data was generated, stored, and used directly in fine-tuning the BERT-based language models. The only modification was to add BERT's "special" [CLS] and [SEP] tokens so the model could process the text.

The OpenAI API provides a method for varying the degree of "aggressiveness" in generating text by adjusting temperature sampling. In this study, we performed tests at temperature values of 0.1, 0.5, and 0.9 to determine if temperature is an important factor in text generation such that it affects model performance.

After GPT-3.5 was used to create artificial student responses to augment small data sets, those augmented data sets were then used to determine if model performance using sentence classification improves or degrades.

3.3 Model Classification

Since we had data sets in two different languages, we chose a multilingual model to compare the use of language when performing fine-tuning. We chose the Microsoft *Multilingual L12 H384* model as a basis for all testing due to its performance gains over the base BERT model and its improved ability for fine-tuning [32]. We fine-tuned it using a combination of original data and augmented data for training. Data was held out from the original data set for testing purposes. A separate BERT model was fine-tuned for each concept and augmentation type to classify the data by adding a single feed-forward layer. This resulted in 28 separate *BERT-based* models that were fine-tuned and evaluated for this study. We used the micro-F_1 metric as the performance measurement. The models were trained and evaluated ten times, with each training iteration using a different seed for the random number generator, which partitions the training and testing instances. The train/test split was 80/20.

3.4 Baseline Evaluation

We evaluated two different baseline models for each concept. The *a priori* model chose the majority classification for each concept. For our *unaugmented* baseline, we applied BERT prototypically without data augmentation or balancing. The baseline performance results are shown in Table 3.

4 Results

Table 3 presents a summary of the results. Each row corresponds to a concept for the English data set, with one row for the French Data set. The leftmost data column shows the percentage of the answers for each concept marked with the majority label. The following two columns present the baseline results. On the

right are the maximum F_1 scores for each concept using either self-augmented or generated data and indicating the augmentation level used to achieve that maximum performance. The highest performance for each data set is shown in bold, indicating which method or data set was used to achieve that score.

Table 3. Performance (micro-F_1) of baseline vs all augmented models

	% Maj.	Baseline		Max Performance		
Concept	Label	*a priori*	Unaug.	Self	GPT-3.5	Aug.
French	73	**0.720**	0.575	0.636	0.612	21x
C1	89	**0.940**	0.735	0.789	0.815	0.6x
C2a	73	0.850	0.757	**0.931**	0.921	8x
C2b	67	0.670	0.547	0.852	**0.874**	55x
C3a	54	0.700	0.532	0.726	**0.815**	55x
C3b	77	0.770	0.684	0.926	**0.952**	55x
C3c	60	0.600	0.568	0.747	**0.832**	89x

Figure 1 illustrates how each of the four augmentation methods affected model performance as more augmentation was used to train the model. The "self" label on the chart indicates the self-augmentation method of creating multiple copies of the original data. The numbers 0.1, 0.5, and 0.9 indicate the temperature setting used on the GPT-3.5 API to provide varied responses, as previously discussed. Note that as augmentation increases, the "self" method peaks and begins to decline in performance with additional training data added, where the augmented models using GPT-3.5 do not drop off as much. This indicates the model is more tolerant of this generated data than continuing to fine-tune on the same small data set, copied multiple times.

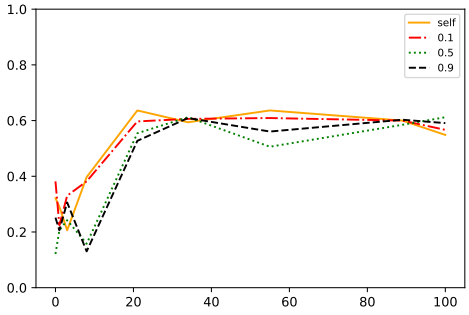

Fig. 1. French Model Performance (micro-F_1) per Augmentation Type. (Note: The x-axis shows the level of augmentation applied from 0x to 100x.)

Figure 2 shows how each of the model's (one for each concept) performances varied with training data using different augmentation types of self, and the other lines indicate each of the three temperatures used to generate text. Note that self-augmentation peaks early while the other types continue improving performance.

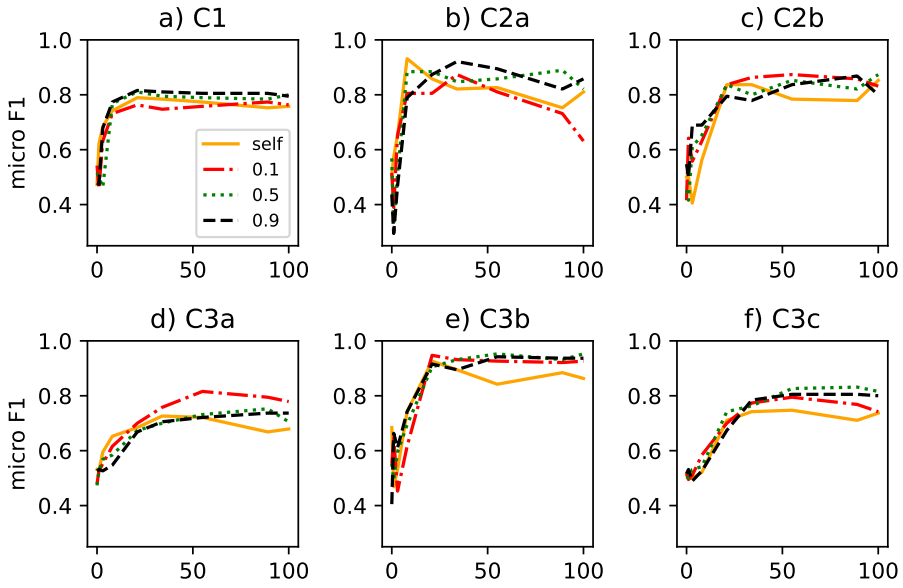

Fig. 2. Rainwater Runoff Model Performance per Augmentation Type. (Note: The x-axis shows the level of augmentation applied from 0x to 100x.)

5 Discussion

Recall **RQ 1**, which asks if artificially-generated responses improve base model performance. This research shows that the augmented model outperformed the unaugmented model in all seven concepts. However, the *a priori* computation which always selects the majority label won on two of the data sets. Our hypothesis **H1** stated that additional augmented data would improve model performance for smaller data sets, and that was shown to be supported by this data. The base model testing without augmentation was always improved upon with augmented data. However, two data sets, C1 and the French Data, were heavily imbalanced. Their entropy values were far from ideal, as shown in Tables 1 and 2. In these cases where entropy is low, guessing the majority label performed better than machine learning models to predict the label for the student response. Determining

augmentation methods that improve performance on low-entropy data sets is an area of further research.

RQ 2 asks if artificially-generated responses outperform self-augmentation when training models for classification. This study shows that the maximum performance was achieved using artificially-generated responses in four out of seven concepts. Our hypothesis **H2** stating that artificially-generated responses will outperform simple replication of existing data as in the self-augmentation method was partially supported. This experiment shows that although simply replicating given data might be produce good performance, generating new examples usually produced the best performance. However, this research also shows that the performance degrades faster when the only data augmentation is from self-augmentation. This needs further research to determine if this is a consistent way to get the model training jump-started before adding other types of augmentation into the mix.

Next, **RQ 3** asks if temperature sampling of the artificially-generated student responses affects model performance. Examining the maximum performance at each augmentation level did not reveal a single winner among the three temperatures used for student response generation. **H3** proposed that augmenting the data with slightly more risky answers, equating to a temperature of 0.5, will provide the best performance in general. In the charts presented in Figures 1 and 2, each data set augmented by the three temperatures varied in performance but were similar to each other. Toward higher values of augmentation, the more risky generation using a temperature of 0.9 continued to increase in performance, indicating reduced overfitting during training. Temperature variation did not significantly alter performance but should be investigated further to see if that is true in general, or only in these specific data sets.

Finally, **RQ 4** ponders if performance ultimately degrades when the model reaches a sufficient level of augmentation. In all models tested, performance peaks and degrades after 55x to 89x augmentation. Table 3 shows the augmentation level at which performance peaked and began to fade. **H4** states that the performance will degrade with additional augmentation, which was supported by all the models tested. **H5** further states that the performance will degrade more slowly with higher temperature and thus more risky generated responses. When examining the performance changes in Figures 1 and 2, the highest temperature, 0.9, rose in performance similar to other temperatures but decreased at a slower pace than the other temperatures, especially at higher augmentation levels. Due to this observation, this hypothesis is supported by the data.

6 Conclusion

This study intended to determine if GPT-3.5 was a viable solution to generate additional data to augment a small data set. We used one multilingual *BERT-based* model, trained it using two different data sets in two languages augmented by two different methods, and compared that result to baseline models against one using self-augmentation and three with GPT-3.5 augmentations. In four out

of seven cases, a model augmented with GPT-3.5 generated responses pushed the performance beyond what could be achieved by other means.

Another objective of this study was to determine if setting the temperature or riskiness in GPT-3.5 response generation would affect performance. Our data shows that while it may not achieve peak results, the higher temperature generated text has more longevity because the models could take on more augmented data and maintain stability than other temperatures or self-augmentation.

These empirical tests show that augmentation methods such as self-augmentation and text generation with GPT-3.5 drastically improve performance over unaugmented models. However, the performance achieved by these models leveled off quickly after augmentation amounts of around fifty times the amount of original data. In addition, two data sets with severely imbalanced data did not improve performance enough to overcome their *a priori* computed values.

Using a higher temperature value when generating data from the GPT-3.5 model did not yield the highest-performing results but came very close. The added benefit of using the higher temperature is that the generated student responses seemed more diverse, allowing the model to prevent overfitting, even at higher augmentation levels.

7 Future Work

In this study, we showed how self-augmentation rises quickly but then levels off and degrades performance. Further research must be done to increase the diversity in generated student responses and research combinations of different augmentation techniques, including GPT-3.5 temperature that might be introduced at different augmentation levels.

References

1. Achieve, Inc: Next Generation Science Standards (2013)
2. Azevedo, R., Johnson, A., Chauncey, A., Burkett, C.: Self-regulated learning with MetaTutor: Advancing the science of learning with metacognitive tools. In: New science of learning, pp. 225–247. Springer (2010)
3. Balkus, S., Yan, D.: Improving short text classification with augmented data using GPT-3. arXiv preprint arXiv:2205.10981 (2022)
4. Basu, S., McElhaney, K.W., Rachmatullah, A., Hutchins, N., Biswas, G., Chiu, J.: Promoting computational thinking through science-engineering integration using computational modeling. In: Proceedings of the 16th International Conference of the Learning Sciences (ICLS) (2022)
5. Bayer, M., Kaufhold, M.A., Buchhold, B., Keller, M., Dallmeyer, J., Reuter, C.: Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International Journal of Machine Learning and Cybernetics pp. 1–16 (2022)
6. Bayer, M., Kaufhold, M.A., Reuter, C.: A survey on data augmentation for text classification. arXiv preprint arXiv:2107.03158 (2021)

7. Biswas, G., Segedy, J.R., Bunchongchit, K.: From design to implementation to practice a learning by teaching system: Betty's brain. International Journal of Artificial Intelligence in Education **26**(1), 350–364 (2016)
8. Britt, M.A., Rouet, J.F., Durik, A.M.: Literacy beyond text comprehension: A theory of purposeful reading. Routledge (2017)
9. van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., Basche, P.: Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. Journal of Educational Psychology **93**(3), 521 (2001)
10. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
11. Chen, J., Tam, D., Raffel, C., Bansal, M., Yang, D.: An empirical survey of data augmentation for limited data learning in nlp. arXiv preprint arXiv:2106.07499 (2021)
12. Cho, Y., Choi, I.: Writing from sources: Does audience matter? Assessing Writing **37**, 25–38 (2018)
13. Cochran, K., Cohn, C., Hastings, P.: Improving NLP model performance on small educational data sets using self-augmentation. In: To appear in the Proceedings of the 15th International Conference on Computer Supported Education (2023)
14. Cochran, K., Cohn, C., Hutchins, N., Biswas, G., Hastings, P.: Improving automated evaluation of formative assessments with text data augmentation. In: International Conference on Artificial Intelligence in Education. pp. 390–401. Springer (2022)
15. 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)
16. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for nlp. arXiv preprint arXiv:2105.03075 (2021)
17. Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in game-based learning environments with word embedding representations of reflection. International Journal of Artificial Intelligence in Education **31** (10 2021). https://doi.org/10.1007/s40593-020-00220-4
18. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)
19. Hastings, P., Hughes, S., Britt, A., Blaum, D., Wallace, P.: Toward automatic inference of causal structure in student essays. In: International Conference on Intelligent Tutoring Systems. pp. 266–271. Springer (2014)
20. Hutchins, N.M., Basu, S., McElhaney, K.W., Chiu, J.L., Fick, S.J., Zhang, N., Biswas, G.: Coherence across conceptual and computational representations of students' scientific models. In: Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021. International Society of the Learning Sciences (2021)
21. Käser, T., Schwartz, D.L.: Modeling and analyzing inquiry strategies in open-ended learning environments. International Journal of Artificial Intelligence in Education **30**(3), 504–535 (2020)
22. Liu, P., Wang, X., Xiang, C., Meng, W.: A survey of text data augmentation. In: 2020 International Conference on Computer Communication and Network Security (CCNS). pp. 191–195. IEEE (2020)

23. Luckin, R., du Boulay, B.: Reflections on the Ecolab and the Zone of Proximal Development. International Journal of Artificial Intelligence in Education **26**(1), 416–430 (2016)
24. McElhaney, K.W., Zhang, N., Basu, S., McBride, E., Biswas, G., Chiu, J.: Using computational modeling to integrate science and engineering curricular activities. In: In M. Gresalfi & IS Horn (Eds.). The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020. vol. 3 (2020)
25. OECD: 21st-Century Readers. PISA, OECD Publishing (2021). https://doi.org/https://doi.org/https://doi.org/10.1787/a83d84cb-en, https://www.oecd-ilibrary.org/content/publication/a83d84cb-en
26. Pilipiszyn, A.: GPT-3 powers the next generation of apps (2021)
27. Quteineh, H., Samothrakis, S., Sutcliffe, R.: Textual data augmentation for efficient active learning on tiny datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7400–7410. Association for Computational Linguistics (2020)
28. Schwartz, R., Stanovsky, G.: On the limitations of dataset balancing: The lost battle against spurious correlations. arXiv preprint arXiv:2204.12708 (2022)
29. Seo, J.W., Jung, H.G., Lee, S.W.: Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning. Neural Networks **138**, 140–149 (2021). https://doi.org/https://doi.org/10.1016/j.neunet.2021.02.007, https://www.sciencedirect.com/science/article/pii/S0893608021000496
30. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
31. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Text data augmentation for deep learning. Journal of big Data **8**(1), 1–34 (2021)
32. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems **33**, 5776–5788 (2020)
33. Winne, P.H., Hadwin, A.F.: nStudy: Tracing and supporting self-regulated learning in the internet. In: International handbook of metacognition and learning technologies, pp. 293–308. Springer (2013)
34. Wu, L., Xie, P., Zhou, J., Zhang, M., Ma, C., Xu, G., Zhang, M.: Self-augmentation for named entity recognition with meta reweighting. arXiv preprint arXiv:2204.11406 (2022)
35. Xia, C., Zhang, C., Zhang, J., Liang, T., Peng, H., Philip, S.Y.: Low-shot learning in natural language processing. In: 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI). pp. 185–189. IEEE (2020)
36. Yogatama, D., d'Autume, C.d.M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al.: Learning and evaluating general linguistic intelligence. arXiv preprint arXiv:1901.11373 (2019)
37. Zhang, N., Biswas, G., McElhaney, K.W., Basu, S., McBride, E., Chiu, J.L.: Studying the interactions between science, engineering, and computational thinking in a learning-by-modeling environment. In: International Conference on Artificial Intelligence in Education. pp. 598–609. Springer (2020)