AI GEN ASSISTMENT: ANALYZING THE EFFECTIVENESS OF GENERATIVE AI DATA AMPLIFICATION FOR DKT

Myong Sung Noh
Department of Elementary Computer Education
Seoul National University of Education
Seocho Jungang-ro 96, Seoul, Korea, 06639
myongsungcs@gmail.com

Ung Hui Cho
Department of Software Engineering
SangMyung University
Sangmyeongdae-gil 31, Cheonan, Korea, 31066
paul9512@gmail.com

ABSTRACT

In edtech, it is one of the important factors to predict the learning level of learners and whether they will solve problems in the future by utilizing a learner tracking system. However, due to the lack of initial data, it is limited to perform such prediction work accurately from the beginning. Also, due to limited data, there are limitations in the performance evaluation method of the model for each data. Therefore, our research team utilized generative artificial intelligence technology to amplify virtual log data in large quantities based on small original data to create a virtual test data set and verify whether it is effective as a learning and evaluation data set for DKT Model. As a result of the experiment, about 100,000 data sets were built from the initial data state of 10,000 data, which confirmed the possibility of learning and evaluating DKT meaningfully.

1 INTRODUCTION:

A learning tracking system called Deep Knowledge Tracing (DKT) utilizes deep learning techniques to predict whether a student will solve a problem in the future based on the logs of the student's solved problems. At this time, various data can be utilized for prediction, and the form of the model has also evolved through the form of RNN, LSTM, and transformer.

However, no matter how good the model and data are, if the absolute amount of initial training data is insufficient, the prediction accuracy of the model will be low, and the process of reliably evaluating the performance of the model will be limited. Our research team sought to solve these problems through generative AI, which has recently emerged as a major technical issue.

We took 10000 data directly extracted from the Assistance 2009 dataset, which is often used in research, and amplified it by about 10 times to about 100,000 virtual data using generative AI, and trained it on the most basic prediction model, which is a bidirectional LSTM, and measured its performance. As a result, we found the possibility that the learning and evaluation performance of DKT models can be as good as the datasets covered in previous studies, and we expect that we can use it to more systematically reveal the process of learning and evaluating DKT models when given cold start problems or other small training datasets in the future.

2 METHODS

First of all, based on Assistment2009, which is a training data that is mainly utilized in research, we extracted 10000 data directly from the research team to prevent overfitting in this dataset and make it representative. In addition, during the extraction process, we extracted only 2 features from the data for ease of learning and smoothness of research, which are 'sequenceid' and 'skillid'. We plan to use more features in future research.

Based on the 10000 corresponding data extracted from Assistance2009, we built a dataset of about 100,000 using GPT-4, assuming the inflow and activity of real users. Our expectation was that this

method would provide data from generative AI that matches the behavior patterns of real learners as closely as possible.

To prove our expectations, we first built a simple DKT model to train and evaluate the dataset. The model was based on a bidirectional LSTM, with a loss function of CrossEntropyLoss, an optimization function of Adam, a learning rate of 0.001, and a dropout to prevent overfitting of the data.

Based on the model, the research team conducted experiments to compare the performance of learning and evaluating the existing research dataset, Assistance 2009, and the virtual dataset created by the research team based on the dataset, and to compare the performance of learning and evaluating the dataset created by the research team.

3 EXPERIMENT

First of all, as mentioned above, the research team prepared two datasets before the experiment: Assistment 2009 and 100,000 data generated based on it.

The ultimate goal of our experiments is to verify whether DKT can learn and evaluate the two datasets similarly when trained on the data created by the model and the Assistment 2009 data through generative AI, and to determine whether generative AI can generate a virtual dataset that reflects the learners' real-world patterns as much as possible.

In our full-scale experiments, we applied 5-fold cross-validation (Kfold 5) to evaluate the reliability and generalization ability of the model. Cross-validation involves dividing the entire dataset into five equal-sized subsets, using four of them as training data and one as validation data, and repeating the process five times. This helps to verify that the model is not overfitted to a particular dataset and can perform consistently on a variety of data.

For training the model, we set the batch size to 64 and trained for a total of 10 epochs. The dataset used for training was based on Assistment 2009, which we fictitiously named "AI Gen Assistment". We selected three main evaluation metrics (AUC, BCE, and F1 score) for training and evaluating the model. These metrics play an important role in evaluating how well the model fits the training data and how well it performs in a real-world test environment.

Table 1: Comparison of Evaluation Metrics

SCORE	**Assistment2009**	**AI gen Assistment**
AUC	0.7101	0.7164
BCE	0.5660	0.6206
F1	0.8075	0.6387

From the above evaluation results, we can observe that the AUC score is similar, but the BCE score is slightly different and the F1 score is significantly different.

4 CONCLUSION

In this study, two virtual datasets, 'Assistment2009' and 'AI Gen Assistment', were trained and evaluated using a deep knowledge tracing (DKT) model. 'AI Gen Assistment' showed a slight advantage in AUC score, but performed worse than 'Assistment2009' in BCE score and F1 score, suggesting that 'AI Gen Assistment' needs improvement in terms of prediction accuracy, class imbalance, and learning pattern recognition. Overall, the 'AI Gen Assistment' dataset showed potential for training DKT models, and these results can serve as an important basis for developing AI-based learning tracking systems in the field of educational data science.

URM STATEMENT

The authors acknowledge that at least one of the lead authors of this paper meets the URM criteria for the ICLR 2024 small paper track

REFERENCES

Markus Endres, Asha Mannarapotta Venugopal, and Tung Son Tran. Synthetic data generation: a comparative study. In *Proceedings of the 26th International Database Engineered Applications Symposium*, pp. 94–102, 2022.

Yingzhou Lu, Huazheng Wang, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

Anastasia Olga, Akash Saini, Gabriela Zapata, Duane Searsmith, Bill Cope, Mary Kalantzis, Vania Castro, Theodora Kourkoulou, John Jones, Rodrigo Abrantes da Silva, et al. Generative ai: Implications and applications for education. *arXiv preprint arXiv:2305.07605*, 2023.

Deepak Kumar Panda and Sanjog Ray. Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review. *Journal of Intelligent Information Systems*, 59(2):341–366, 2022.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

Xiangyu Song, Jianxin Li, Taotao Cai, Shuiqiao Yang, Tingting Yang, and Chengfei Liu. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, 258:110036, 2022.

Maarten van der Velde, Florian Sense, Jelmer Borst, and Hedderik van Rijn. Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. *Computational Brain & Behavior*, 4:231–249, 2021.

Mustafa Yağcı. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11, 2022.

Yupei Zhang, Yue Yun, Rui An, Jiaqi Cui, Huan Dai, and Xuequn Shang. Educational data mining techniques for student performance prediction: method review and comparison analysis. *Frontiers in psychology*, 12:698490, 2021.

Lu et al. (2023) van der Velde et al. (2021) Zhang et al. (2021) Yağcı (2022) Piech et al. (2015) Song et al. (2022) Panda & Ray (2022) Olga et al. (2023) Endres et al. (2022)

A APPENDIX

A.1 DATASET COMPARISON SUMMARY

Table 2: Dataset Comparison - Part 1

Dataset	Unique problem IDs	Unique user IDs	Average achievement	User-wise duplicate data average
corrected_augmented_10k	18,580	4,136	48.52%	0.57
10k	4,706	4,136	48.52%	0.04
assist2009	26,688	4,217	62.23%	0.95

Introduction

This section offers a succinct comparison between the assist2009 dataset, consisting of genuine student problem-solving logs, and its subsets 10k and the GPT-generated corrected_augmented_10k.

Data Origin

Table 3: Dataset Comparison - Part 2

Dataset	Unique skill IDs	Unique sequence IDs	User-wise unique skill IDs average	User-wise unique sequence IDs average
corrected_augmented_10k	114	1,756	1.87	12.96
10k	114	516	1.87	1.94
assist2009	123	677	9.96	14.2

- assist2009: Original data capturing diverse student interactions with 26,688 unique problems and 4,217 unique users.
- 10k: A diverse subset from assist2009, retaining all users but limiting to 10,000 data points for broader problem coverage.
- corrected_augmented_10k: GPT-created dataset, expanding on 10k to reproduce the extensive problem variety found in assist2009.

Unique Identifiers Analysis

- The corrected_augmented_10k mirrors assist2009 in unique problem IDs, indicating a successful emulation of problem variety.
- Unique user IDs remain unchanged from 10k to corrected_augmented_10k, ensuring user representativeness.
- Skill and sequence ID distributions suggest corrected_augmented_10k aligns closely with assist2009, which supports data integrity.

Achievement and Duplication

- assist2009 shows a higher average achievement rate, reflecting richer interaction data.
- Duplicate data averages point to more repeat interactions in assist2009, aligning with its larger size and diversity.

Conclusion

The comparison illustrates that corrected_augmented_10k effectively simulates the original dataset's complexity, with comparable diversity in problems and user engagement. These insights validate the use of synthetic data for AI model training, offering an expanded platform while noting the necessity of considering the unique attributes of genuine data in model application and interpretation.

A.2 Data Extraction Methodology Summary

Objective

This appendix delves into the methodologies employed for data extraction across three distinct datasets used in the study: assist2009, 10k, and corrected_augmented_10k.

Original Dataset (assist2009)

The assist2009 dataset is a repository of genuine student engagement logs, detailing interactions with various problems. It stands as the foundational dataset, encompassing a wide array of user interactions across 26,688 unique problem IDs and 4,217 unique user IDs.

Subset Creation (10k)

From the assist2009 dataset, a subset labeled 10k was meticulously curated to ensure a comprehensive representation while maintaining manageability. The selection process prioritized the retention of all unique user IDs to preserve user diversity. Simultaneously, it aimed to encapsulate a broad spectrum of problem interactions, capping the dataset at 10,000 data points to strike a balance between diversity and dataset size.

Synthetic Data Generation (corrected_augmented_10k)

The creation of the corrected_augmented_10k dataset was driven by the need to amplify the dataset size while maintaining the diversity and complexity observed in the original logs. Utilizing GPT, a generative model was tasked to synthesize data based on patterns learned from the 10k subset. The generation process was fine-tuned to match the original assist2009 dataset in terms of unique problem IDs, thereby extending the problem space without compromising the distribution of user interactions.

Methodological Considerations

The extraction for 10k required balancing the breadth of problem coverage against the depth of user interactions. For corrected_augmented_10k, the generative process involved algorithmic mimicry of the 10k problem-solving patterns, necessitating a careful calibration to ensure data fidelity. While specifics of the GPT's generation strategy are abstracted, the resulting dataset's alignment with assist2009's diversity benchmarks underscores the efficacy of the approach.

Conclusion

The extraction and generation methodologies for the datasets articulate a deliberate effort to craft data landscapes that are not only reflective of authentic user problem-solving behavior but also conducive to robust AI model training. corrected_augmented_10k particularly embodies an innovative stride in dataset augmentation, leveraging AI to fulfill the dual objectives of scale and representativeness.