GPT-3.5 for Data Augmentation in Automatic Essay Scoring: A Preliminary Analysis

Ruan Carvalho[1], Péricles B.C. Miranda[1], Hilário T.A. Oliveira[2], Cleon Xavier[3], Luiz Rodrigues[4], Newarney T. Costa[3], Rafael Ferreira Mello[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

[2]Instituto Federal do Espírito Santo

[3]Instituto Federal Goiano

[4]Universidade Federal de Tecnologia do Paraná

{rafael.mello,pericles.miranda}@ufrpe.br

Abstract. *Machine learning models are susceptible to the dataset used during its training. Dealing with limited or imbalanced datasets is challenging, and a commonly adopted approach to mitigate this limitation is data augmentation. For example, expanding the training set in a computer vision problem may involve rotation and resizing images; however, this task is more complex when dealing with textual data. This work investigates the use of GPT-3.5 for data augmentation in a dataset of argumentative essay texts from the National High School Exam (ENEM), which is used as a selection criterion for entry into public universities in Brazil. More specifically, we adopted traditional Natural Language Processing (NLP) techniques for essay scoring and compared the results with and without the data augmentation. Our results show that the long argumentative essays generated by GPT in the data augmentation process did not improve the performance of NLP models. Moreover, GPT could not adequately classify its synthetic data, suggesting poor quality of the generated data, and did not outperform NLP models in classifying real data.*

1. Introduction

The National High School Examination (ENEM) is one of Brazil's primary assessments to gauge students' educational proficiency after their secondary education. It is through ENEM that numerous students gain access to higher education in public and private institutions[1]. ENEM encompasses a series of questions, including multiple-choice and open-ended questions, as well as an essay (ENEM essay), where exam administrators present a theme, a contentious issue, and motivational texts to demonstrate the presence of the challenging issue within society [ENEM 2022].

The ENEM essay is a dissertative-argumentative text in which students usually propose a solution to the societal issue presented within 30 lines [ENEM 2022]. The evaluation of the essay is based on five criteria (or competencies): C1) Demonstrating mastery of the formal written mode of the Brazilian Portuguese language; C2) Comprehending the essay prompt and applying concepts from various fields of knowledge to

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

DOI: 10.5753/sbie.2025.12536

elaborate on the topic; C3) Selecting, correlating, organizing, and interpreting information, facts, opinions, and arguments in defense of a particular standpoint; C4) Exhibiting an awareness of the linguistic mechanisms necessary for constructing the argumentation; and C5) Devising an intervention proposal for the addressed problem while upholding human rights. Each competency is scored from 0 to 200, with a total essay score ranging from 0 to 1000. Two assessors evaluate each essay and, in case of significant score differences, a third assessor is involved to reach a consensus [2].

The manual grading process, although necessary, is widely acknowledged for its significant drawbacks, particularly concerning the fatigue experienced by evaluators due to its repetitive nature [Burrows et al. 2015]. Furthermore, due to its human-centric nature, the grading process is susceptible to numerous inconsistencies and biases, which may result in unreliable assessments [Ferreira Mello et al. 2022]. A promising solution to this issue is the deployment of intelligent computational systems that can automate the grading process, improving efficiency and ensuring more consistent and impartial outcomes [Oliveira et al. 2023a, Ferreira Mello et al. 2022]. Natural Language Processing (NLP) techniques have been developed to address this need by automatically processing essays and automatically generating corresponding grades using Machine Learning (ML) models [Marinho et al. 2022b, Carvalho et al. 2024, Oliveira et al. 2023a, Galhardi et al. 2024]. This task is referred to in the literature as Automated Essay Scoring (AES) [Dikli 2006].

A challenge faced in this domain is data imbalance. Generally, PT-BR corpora used in most studies have a major class with 18 times more instances than the minor one [Costa et al. 2020, Oliveira et al. 2023a]. For instance, high-scoring essays are not as common as low-scoring ones [de Lima et al. 2023]. While related work has explored data augmentation in this task, it has either focused on essays in English [Dai et al. 2023, Cochran et al. 2023] or been limited to traditional techniques such as SMOTE [Oliveira et al. 2022]. In contrast, recent advancements in Large Language Models (LLMs) have raised questions about their utility for data augmentation. However, to the best of our knowledge, previous research has not investigated LLMs' contribution to data augmentation for AES in ENEM essays.

In an effort to contribute to the field of AES in PT-BR, particularly regarding data imbalance, this study performs an experimental analysis of strategies to estimate Competency 3 (C3) of the ENEM exam considering the extended Essay-BR dataset [Marinho et al. 2022a], which comprises 6,579 essays from the ENEM. To address the data imbalance issue, the present study employed data augmentation, an approach commonly used in the literature [Quteineh et al. 2020, Bayer et al. 2022]. In this approach, the LLM GPT-3.5 was utilized to generate new essays based on the original essay topics. The goal of the study was to determine if data augmentation using GPT-3.5 leads to better ML models.

2. Background

AES refers to the process of automatically assigning grades to an essay based on pre-established correction criteria. A common approach in the literature involves the use of

[2] https://www.gov.br/inep/pt-br/assuntos/noticias/enem/conheca-o-processo-de-correcao-das-redacoes

NLP techniques and ML algorithms trained on annotated corpora with human scores in one or multiple competencies [Chassab et al. 2021]. Feature-based approaches and, more recently, models based on deep neural networks have been used to develop new AES solutions. Furthermore, a new path that has been investigated is the use of hybrid approaches, which combine handcrafted features and representations extracted from neural models, such as the Bidirectional Encoder Representations from Transformers (BERT) [Li et al. 2023]. The proliferation of LLMs has opened new avenues for enhancing AES across its different stages, demonstrating promising results and further improving the accuracy and consistency of scoring systems [Chassab et al. 2021, Xiao et al. 2024].

A significant challenge in AES lies in the limited availability of human-annotated essays for training ML models. However, creating such data is both costly and time-consuming, posing a barrier to the development of robust models [Park et al. 2022]. This issue becomes even more pronounced when considering the scarcity of datasets in languages other than English, complicating efforts to create non-English and multilingual AES systems [Costa et al. 2020, Bai and Stede 2022]. Furthermore, the lack of balanced and representative data on essay scores undermines the robustness of AES models, introducing bias into the scoring process, as the system tends to be influenced by the most represented data samples or be affected by disturbances introduced at the word or phrase level [Philip and Tashu 2024]. To address this issue, data augmentation is a widely used technique. However, there is still a limited number of studies exploring augmentation techniques specifically at the phrase level [Philip and Tashu 2024].

Given the challenges posed by limited and unbalanced datasets in the AES domain, exploring LLMs for data augmentation represents a highly relevant research direction [Xiao et al. 2024]. Data scarcity not only limits the generalizability of models but also introduces biases in scoring, disproportionately affecting underrepresented samples [Philip and Tashu 2024]. LLMs offer a promising solution to generate synthetic data that closely mimics real-world essays, thereby expanding the dataset and significantly improving the balance across scoring categories. Moreover, LLMs can enhance the quality of augmented data by producing linguistically diverse and contextually rich examples, which are crucial for training more accurate and robust AES models [Xiao et al. 2024]. However, research on LLM-based data augmentation is still in its early stages, as discussed next.

2.1. LLM-based Data Augmentation

Early research relied on fine-tuning LLMs for data generation. Quteineh et al. [Quteineh et al. 2020] utilized a small portion of the training set to fine-tune the GPT-2 model, enabling it to generate new sentences. The authors compared this approach, applied to English texts, with Non-Guided Data Generation (NGDG) and achieved a 5% performance improvement.

In [Bayer et al. 2022], fine-tuning was also performed on GPT-2 by adding tokens before each text to indicate the data class, and the model's temperature parameter was adjusted to introduce uncertainty in generating new sentences. Subsequently, GPT-2 generated new texts, and a filter was applied to increase the probability that the synthetic data maintains the original labels. This methodology was applied to 11 datasets (all in English), achieving improvements of two to four points in F1-score in some datasets.

Recent studies have explored the advanced capabilities of LLMs, such as GPT-3.5. For instance, Cochran et al. [Cochran et al. 2023] used the GPT-3.5 model (text-curie-001) with the command "paraphrase this sentence" and an instance of the real dataset for data augmentation. Three values of the model's "temperature" parameter (i.e., determines whether the output is more random or predictable) were tested to investigate whether it affected the performance of the BERT model as a final classifier. These approaches were compared with a baseline (BERT model without data augmentation), self-augmentation, and a priori model (always choosing the majority class). The data augmentation approaches outperformed the baseline in all seven datasets studied, although the GPT-based method outperformed the other methods in only four datasets. Authors also observed that higher "temperature" values generated sentences with greater diversity, although this did not alter the overall performance of models trained with synthetic data from GPT-3.5.

Similarly, Dai et al. [Dai et al. 2023] proposed AugGPT, which uses ChatGPT to rewrite each sentence in the training dataset and produce several new sentences while preserving the semantics of the original data. The results of applying AugGPT were compared with those of 21 other data augmentation methods on three datasets, and the proposed method was associated with better accuracy in all scenarios. Hence, these findings provide promising evidence of the potential of using advanced LLMs for data augmentation in text-based tasks. Nevertheless, no previous research has explored the capabilities of these models in augmenting data for AES in the context of the ENEM exam.

2.2. Research Questions

Inspired by previous research, we leveraged GPT-3.5 to generate essays and investigate whether synthetic data can lead to improved models for estimating scores in C3 of the ENEM, specifically to address the data imbalance issues. Beyond the language aspect, this study distinguishes itself from others due to the length of the texts utilized. On average, each essay comprises 12 sentences and 290 words, making them considerably lengthy compared to related work. For example, in [Bayer et al. 2022], texts were considered short if they contained up to 280 characters. Thus, our first research question was:

> **RESEARCH QUESTION 1 (RQ1):**
> *Does the synthetic data produced by the GPT-3.5 improve the machine learning model's performance in estimating the C3 competence on essays?*

Furthermore, it is important to ensure that synthetically generated data is of high quality. Alternatively, given that real-world data also suffers from biases, synthetic data should at least be of comparable quality when contrasted with real-world data. In that regard, synthetic data should be at least representative enough to create ML models with performance comparable to those trained on real-world data. Therefore, our second research question was:

> **RESEARCH QUESTION 2 (RQ2):**
> *How do GPT-generated synthetic data compare to real-world data in terms of fitting AES models?*

3. Methodology

3.1. Original Dataset

The present study used the Extended Essay-BR dataset, a corpus comprising 6,579 argumentative essays distributed across 151 different topics [Marinho et al. 2022a]. The corpus was created to fill the resource gap for developing alternative methods for the automatic assessment of essays in Portuguese. The dataset consisted of multiple essays written by Brazilian high school students on an online platform and evaluated by experts in five competencies. The evaluation process follows similar criteria adopted in the ENEM [Marinho et al. 2022a].

A significant challenge in the Extended Essay-BR is data imbalance. It is common for scores at the extremes to have few examples. Table 1 displays the distribution of essays by C3 scores, which is the focus of this research. Scores 0, 40, and 200 had fewer than two hundred examples, while the majority score had over 3,000 examples.

Table 1. Distribution of essay grades considering C3.

Grade	0	40	80	120	160	200
Number of essay	185	164	1,601	3,051	1,374	190

3.2. Automatic Essay Scoring Approach

In the present study, we used a feature-based approach to estimate C3 grades. We adopted traditional features generated using NLP tools and embedding representations extracted using the BERT model [Oliveira et al. 2023a, Galhardi et al. 2024, Carvalho et al. 2024]. A set of 236 features was computed for each essay using the Portuguese versions of Coh-Metrix [Camelo et al. 2020] and Linguistic Inquiry Word Count (LIWC) [Carvalho et al. 2019] tools. Additionally, BERTimbau (neuralmind/bert-base-portuguese-cased) [Souza et al. 2020] was adopted for extracting 768 embeddings from both the raw essay text and the motivating text. Thus, a total of 1,772 different features were extracted from each essay. We developed this methodology based on previous work in the literature [Carvalho et al. 2024].

For machine learning algorithms, we selected the most used and accurate in previous works for AES for Portuguese documents [de Lima et al. 2023, Costa et al. 2020, Ferreira Mello et al. 2022], which included LGBM [3], and XGBoost [4], with both classification and regression versions. The algorithms adopted were configured with their default parameterization, as defined in their respective libraries. These models were trained using features extracted from the text and embeddings obtained from the text and the prompt using the BERT model [Oliveira et al. 2023a].

We trained these models to estimate the C3 score using both classification and regression approaches. Originally, the scores were discrete (classes: 0, 40, 80, 120, 160, 200), with an associated order. For the classification experiments, we considered the classes [0, 1, 2, 3, 4, 5]. However, by modeling the problem this way, the model would not be able to preserve the order among classes. Nevertheless, the advantage was that the model remained faithful to the original problem, providing one of the six possible outputs for each input essay [Oliveira et al. 2023a].

An alternative was to adapt the problem for regression. In this format, after normalization, the inputs would have only six possible values (0, 0.2, 0.4, 0.6, 0.8, 1.0). However, the algorithm could evaluate new inputs (test sets) with intermediate scores

[3] `https://github.com/Microsoft/LightGBM/`

[4] `https://github.com/dmlc/xgboost/`

(e.g., 0.43) during training. This way, the ordering would be preserved (e.g., 0.63 is greater than 0.41), and intermediate values were possible, allowing the algorithm to express its "uncertainty" through an intermediate score. Nonetheless, the final score had to fall into one of the six categories: [0, 1, 2, 3, 4, 5]. For this purpose, the output of the regressor is multiplied by 5 (the value of the highest class) and then rounded to the nearest integer. If the original output of the regressor were less than 0 or greater than 5, the final value would be mapped to 0 or 5, respectively [Oliveira et al. 2023b].

The following metrics were chosen to evaluate the algorithms' performances: Accuracy, F1-score, Cohen's Kappa, Square Kappa, Pearson Correlation, and Root Mean Square Error (RMSE). We chose these metrics informed by related work [Oliveira et al. 2023a, Carvalho et al. 2024], aiming to present a comprehensive evaluation of the AES performance.

3.3. LLM-based Data Augmentation

We conducted three experiments to investigate how LLM-based data augmentation impacts AES in the context of predicting C3 scores in ENEM essays.

To answer the first research question, initially, we performed a preliminary analysis of the GPT-3.5[5] contribution to data augmentation. In the context of ENEM, a motivational text provides data for students to discuss a social problem and, at the end of the essay, present a proposal for intervention, that is, a possible viable solution for the case itself. We had initial conversations with GPT 3.5, who demonstrated a solid understanding of the ENEM essay structure, the five competencies assessed, and the scoring criteria (0, 40, 80, 120, 160, 200) for each.

From the OpenAI library, we used GPT-3.5 to generate a sample of synthetic essays for the minority classes of the Extended Essay-BR dataset, leading to 5 essays by theme (755) for the zero score, 4 essays by theme (604) for score 40, and 6 essays by theme (906) for score 200. This approach resulted in the generation of 2265 new essays, which was accomplished by prompting GPT-3.5 with the text described in Table 2.

Table 2. Prompt used on GPT-3.5.

Prompt	role = system	role = user
Portuguese *(original)*	Você é um aluno prestes a concluir o ensino médio no Brasil.	Crie uma redação nota X na Competência 3 do ENEM com cerca de 200 palavras contendo no máximo 4 parágrafos e com base nos textos motivadores a seguir: TEXTO_MOTIVADOR
English *(translation)*	You are a student who is completing high school in Brazil.	Write an essay scoring X in Competency 3 of the ENEM with approximately 200 words, containing a maximum of 4 paragraphs, and based on the following motivating texts: MOTIVATING_TEXT

[5] At the time the synthetic data was generated to assess whether the use of data augmentation with LLMs would improve the performance of [Carvalho et al. 2024], the GPT 4 model was not yet available.

Note that even after including the 2,265 essays of the minority classes, these data instances remained insufficient to achieve a balanced dataset. It was performed to provide preliminary insight into the contribution of GPT-3.5 as a data augmentation strategy, following literature suggestions that adding too many instances to balance a dataset could harm, rather than improve, AES performance [Bayer et al. 2022]. To alleviate the imbalance in this scenario, we utilized the *compute_class_weight* function from the *scikit-learn* library, which assigns weights based on the number of examples, guiding the model to focus on the minority classes [Quteineh et al. 2020].

Then, we evaluated this approach with a 5-fold cross-validation that separated synthetic and real data [Cochran et al. 2023, Park et al. 2022]. In each fold, synthetic data from the minority classes (0, 40, and 200) were incorporated alongside real data for model training. Nonetheless, the essays in the test set consisted of original data. Additionally, the machine learning algorithms were also compared without adding synthetic data to the original dataset. Hence, we ensured that our test set was a reliable representation of real, well-labeled data while we sought to understand the effect of data augmentation on the models' performance.

Moreover, we also employed GPT-3.5 to generate a balanced dataset, where each class has an equal number of instances. Here, our approach differed from that of the initial analysis in two points. First, we utilized the **GPT-3.5-turbo-0613** model, which should achieve better results. Second, for each of the 151 essay themes available on the Extended Essay-BR dataset, 16 synthetic essays were created for each of its six classes. Consequently, a total of 14,496 essays were generated (2,416 per level). Then, we built upon this synthetic dataset to perform two analyzes, both based on a five-fold cross-validation strategy where 80% of the real data was utilized for training and the remaining 20% for testing. First, synthetic data were added at each stage of the cross-validation training to ensure an equal distribution for each class, resulting in the following additions of synthetic data: 2,292 for level 0, 2,309 for 40, 1,159 for 80, none for 120, 1,340 for 160, and 2,288 for 200. Second, we completely merged the original and the synthetic datasets. Therefore, these analyses enabled us to understand how different forms of integrating synthetic data into the training procedure contribute to AES.

To answer research question 2, we investigated the quality of the synthetic essays generated by GPT-3.5-turbo-0613 in our experiments. For this, we trained models using fully balanced synthetic data (14,496 synthetic essays with 2,416 per class) and tested the performance of these models on real data. If the models trained on synthetic data can generalize to real data or at least yield performance comparable to that of previous experiments, it suggests that the synthetic data is a valid approximation of real-world examples. Therefore, this experiment helps us to understand the validity of the synthetic data generated with this study's approach.

4. Results

4.1. RQ1: Does the synthetic data improve the machine learning model?

Table 3 shows the results of the initial analysis, where we highlight the top two values for each metric. Note that (c) and (r) refer to the algorithms in classification and regression modes, respectively. The first column indicates whether the augmentation was applied or not. Overall, modeling the AES problem as a classification task and not using data augmentation yielded better results in most evaluation metrics of the experiments.

Table 3. Results on experiments with class weights and data augmentation for minority classes.

Augmentation	Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
Yes	LGBM(c)	0.5994	0.3890	0.4657	0.5391	0.8627	0.5437
Yes	XGB(c)	0.5998	0.3810	0.4573	0.5283	0.8596	0.5359
No	LGBM(c)	**0.6050**	**0.4111**	**0.4814**	**0.5744**	**0.7986**	**0.5894**
No	XGB(c)	**0.6039**	**0.4271**	**0.4747**	0.5584	0.8294	0.5673
Yes	LGBM(r)	0.5279	0.3196	0.4220	0.5517	0.8439	0.5591
Yes	XGB(r)	0.5046	0.3301	0.3957	0.5243	0.8924	0.5269
No	LGBM(r)	0.5464	0.3593	0.4569	**0.5883**	**0.8228**	**0.5918**
No	XGB(r)	0.5021	0.3236	0.3975	0.5307	0.8862	0.5330

Figure 1 presents the confusion matrices of the LGBM classifier, which achieved the best results, indicating that the use of synthetic data in classes 0, 1, and 5 (scores 0, 40, and 200, respectively) did not enhance the model's performance in these groups. Using data augmentation improved the estimation accuracy rate of essays with 0, 40, and 80 grades, but this improvement was very slight. On the other hand, the accuracy rate in predicting essays with scores of 120, 160, and 200 decreased. Thus, these results demonstrate that including only a few samples of synthetic data for the minority classes did not improve the model's predictive performance or its ability to handle the minority classes.

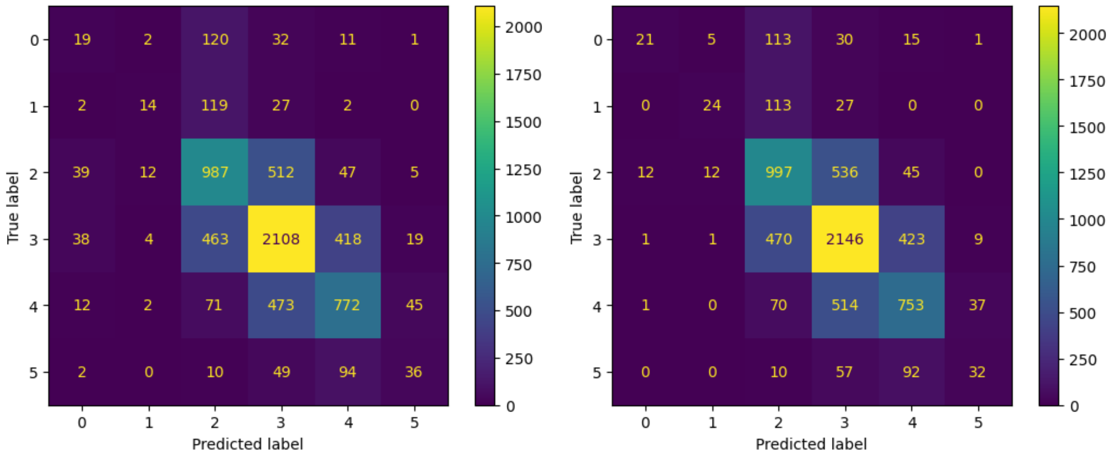

Figure 1. Confusion matrix from the best model from LGBM classifier, where the first and second columns present results with and without data augmentation, respectively.

The results achieved by the LGBM and XGB after adding samples until the dataset is balanced are presented in Table 4, while Table 5 shows the results when the entire balanced set of synthetic data was included in the training set. In both cases, the results did not surpass those achieved by models trained without data augmentation in any of the six metrics adopted in this study. This finding demonstrates that neither fully balancing

Table 4. Synthetic data for balancing and classifying real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.5951**	**0.3454**	**0.4335**	**0.496**	**0.8620**	**0.5103**
XGB(c)	**0.5806**	**0.3430**	**0.4121**	0.4687	0.8890	0.4810
LGBM(r)	0.5444	0.2986	0.3639	0.4603	**0.8509**	**0.4978**
XGB(r)	0.5068	0.3013	0.3539	**0.469**	0.8934	0.4811

the class distribution with synthetic data nor using the entire balanced set of synthetic data resulted in performance improvements.

Table 5. All synthetic data used in the data augmentation to classify real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.5927**	**0.3426**	**0.4179**	**0.4821**	**0.8503**	**0.5064**
XGB(c)	**0.5837**	**0.3383**	**0.4066**	**0.4767**	0.8545	**0.5007**
LGBM(r)	0.5420	0.2942	0.3561	0.4556	**0.8478**	0.4980
XGB(r)	0.5030	0.2918	0.3446	0.4636	0.8911	0.4781

4.2. RQ2: Training the model with only synthetic data

Table 6 presents the performances of the models fully trained in the synthetic dataset generated with GPT-3.5 and evaluated in the real-world data. The table demonstrates that all six evaluation metrics significantly deteriorated, with no configuration achieving 20% accuracy. This finding raises concerns about the quality of the synthetic data, as the LLM-generated samples were unable to represent patterns at least as well as those from the real dataset.

Table 6. Results achieved by models trained on the fully-balanced synthetic data to classify real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	0.1145	0.0993	**0.0925**	0.1623	2.3789	0.3237
XGB(c)	0.1032	0.0944	0.0888	0.1642	2.3464	0.3187
LGBM(r)	**0.1995**	**0.1178**	**0.1075**	**0.2373**	**1.7036**	**0.4175**
XGB(r)	**0.1622**	**0.1154**	0.0761	**0.1776**	**1.8864**	**0.3412**

5. Discussion

This paper focuses on the findings and implications of GPT-3.5 for data augmentation in the context of AES for Brazilian Portuguese language essays, particularly for the ENEM exam. The experimental analysis showed mixed results, highlighting the potential and limitations of GPT-3.5-generated synthetic data.

Firstly, one key observation is that the synthetic data generated using GPT-3.5 did not significantly improve the performance of AES models, especially when predicting scores for essays in minority classes (i.e., those with lower or higher scores). Despite data augmentation being widely regarded as a solution to dataset imbalance, the models trained with GPT-3.5-augmented data did not outperform those trained with real-world data alone, as previous works [Dai et al. 2023, Cochran et al. 2023].

Secondly, the study shows that models trained solely on GPT-3.5-generated essays performed poorly on real data, with significant drops in accuracy and F1 scores, highlighting the lack of complexity and authenticity in synthetic texts. This result is consistent with prior work suggesting that synthetic data is more effective when combined with real-world data [Bayer et al. 2022, Quteineh et al. 2020]. Moreover, adding synthetic data sometimes reduced performance, indicating that LLM-based balancing may introduce noise instead of meaningful variation. Thus, while LLMs have potential for data generation, their integration into training pipelines requires careful consideration.

This study provides three key practical insights for AES in education. First, further research is required on LLMs for data augmentation, as GPT-3.5-generated essays did not improve AES performance, urging caution in their use. Second, large-scale deployment of AES, particularly in high-stakes exams such as ENEM, still demands human oversight and validation [Chassab et al. 2021, Park et al. 2022]. Finally, the findings reinforce the importance of data quality over quantity: synthetic data lacked the diversity and complexity of human language [Ferreira-Mello et al. 2019, Bayer et al. 2022], indicating that stakeholders should prioritize high-quality, well-labeled, real-world datasets over synthetic expansion [Xiao et al. 2024].

6. Limitations and future work

This study provides initial results on GPT-3.5 for data augmentation in AES but points to the need for further research. Two main directions are highlighted: first, the quality of synthetic essays remains a limitation, as GPT-3.5 outputs lacked the complexity and authenticity of real essays, resulting in weak model performance. Future work could test newer LLMs [Chang et al. 2024] and, importantly, explore improved prompt design, since well-crafted prompts can significantly enhance performance [White et al. 2023]. Second, the study's scope is limited to Competency 3 (C3) of the ENEM exam in Brazilian Portuguese. While valid, this focus restricts generalizability, and future studies should investigate whether synthetic data yields better outcomes across other competencies, assessment criteria, and languages.

References

Bai, X. and Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, pages 1–39.

Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C. (2022). Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics*, 14(1):135–150.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Camelo, R., Justino, S., and de Mello, R. F. L. (2020). Coh-metrix pt-br: Uma api web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, pages 179–186. SBC.

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., and Guedes, G. P. (2019). Evaluating the brazilian portuguese version of the 2015 liwc lexicon with sentiment analysis in social networks. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 24–34. SBC.

Carvalho, R., Lins, L. F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I. I., Isotani, S., and Mello, R. F. (2024). Exploring nlp and embedding for automatic essay scoring in the portuguese. In *International Conference on Artificial Intelligence in Education*, pages 228–233. Springer.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Chassab, R. H., Zakaria, L. Q., and Tiun, S. (2021). Automatic essay scoring: A review on the feature analysis techniques. *International Journal of Advanced Computer Science and Applications*, 12(10).

Cochran, K., Cohn, C., Rouet, J. F., and Hastings, P. (2023). Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. In Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O. C., and Dimitrova, V., editors, *Artificial Intelligence in Education*, pages 217–228. Springer Nature Switzerland.

Costa, L., Oliveira, E., and Júnior, A. C. (2020). Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412, Porto Alegre, RS, Brasil. SBC.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., and Li, X. (2023). Auggpt: Leveraging chatgpt for text data augmentation.

de Lima, T. B., da Silva, I. L. A., Freitas, E. L. S. X., and Mello, R. F. (2023). Avaliaçao automática de redaçao: Uma revisao sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).

ENEM (2022). *A redação no Enem 2022: cartilha do participante*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., and Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 404–414.

Galhardi, L., Herculano, M. F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I. I., Isotani, S., and Mello, R. F. (2024). Contextual features for automatic essay scoring in portuguese. In *International Conference on Artificial Intelligence in Education*, pages 270–282. Springer.

Li, F., Xi, X., Cui, Z., Li, D., and Zeng, W. (2023). Automatic essay scoring method based on multi-scale features. *Applied Sciences*, 13(11):6775.

Marinho, J., Anchiêta, R., and Moura, R. (2022a). Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1):65–76.

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022b). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023a). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Oliveira, H., Mello, R. F., Miranda, P., Alexandre, B., Cordeiro, T., Bittencourt, I. I., and Isotani, S. (2023b). Classificaçao ou regressao? avaliando coesao textual em redaçoes no contexto do enem. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1226–1237. SBC.

Oliveira, H., Miranda, P., Isotani, S., Santos, J., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Estimando coesão textual em redações no contexto do enem utilizando modelos de aprendizado de máquina. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 883–894. SBC.

Park, Y.-H., Choi, Y.-S., Park, C.-Y., and Lee, K.-J. (2022). Essaygan: Essay data augmentation based on generative adversarial networks for automated essay scoring. *Applied Sciences*, 12(12):5803.

Philip, H. and Tashu, T. M. (2024). Phrase-level adversarial training for mitigating bias in neural network-based automatic essay scoring. *arXiv preprint arXiv:2409.04795*.

Quteineh, H., Samothrakis, S., and Sutcliffe, R. (2020). Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7400–7410. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., and Fu, Q. (2024). Human-ai collaborative essay scoring: A dual-process framework with llms.