Comparing Text Augmentation by GPT-3.5 and Llama3 for Evaluating Student Responses

Keith Cochran[1*], Clayton Cohn[2], Jean Francois Rouet[3], Peter Hastings[1*]

[1]School of Computing, DePaul University, 243 South Wabash Avenue, Chicago, Illinois, 60604, USA.
[2]Department of Computer Science, Vanderbilt University, Nashville, Tennessee, 37240, USA.
[3]Université de Poitiers, 86073, Poitiers Cedex 9, France.

*Corresponding author(s). E-mail(s): kcochr11@depaul.edu; phasting@depaul.edu;
Contributing authors: clayton.a.cohn@vanderbilt.edu; jean.francois.rouet@univ-poitiers.fr;

Abstract

Writing is a critical educational task because it encompasses so many skills necessary for the modern world, including vocabulary and grammar acquisition, critical thinking, adapting to different audiences, and determining how best to communicate one's ideas. However, written assignments are notoriously time-consuming for teachers to grade, and timely feedback is critical for students' learning. Automated evaluation can provide quick student feedback while easing the manual evaluation burden for teachers. Current machine learning-based methods of evaluating student textual responses have met with varying degrees of success. One main challenge in training these models is the scarcity of student-generated data. Large volumes of training data are needed to create accurate models, and few educational tasks are large enough. To overcome this data scarcity issue, text augmentation techniques have been used to balance and expand the data set so that classification models can be trained with higher accuracy, providing more useful feedback for teachers and students. This paper examines the performance of text augmentation using two Large Language Models (LLMs) to provide supplemental texts for training models for classifying student answers in English and French educational tasks. Our results show that text generation can dramatically improve model performance on small data sets over simple self-augmentation, especially when the LLM is set to generate more varied responses.

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 (Geden et al., 2021; Käser & Schwartz, 2020; Luckin & du Boulay, 2016). By allowing students to choose how they approach and tackle open-ended tasks (Zhang et al., 2020), 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 (Azevedo, Johnson, Chauncey, & Burkett, 2010; Biswas, Segedy, & Bunchongchit, 2016; Winne & Hadwin, 2013).

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 (Achieve, Inc, 2013), 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 (OECD, 2021). Additionally, research in discourse psychology suggests that students' reading strategies are shaped by their assigned reading task and other contextual dimensions (Britt, Rouet, & Durik, 2017). 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 (van den Broek, Tzeng, Risden, Trabasso, & Basche, 2001). Similarly, students' writing is influenced by their perception of the audience (Cho & Choi, 2018).

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 (Hastings, Hughes, Britt, Blaum, & Wallace, 2014). Transformer-based models like BERT have revolutionized the NLP field due to their pre-training on large data sets such as Wikipedia and BookCorpus (Devlin, Chang, Lee, & Toutanova, 2018), which gives 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 data set. However, these models still require sufficient training data from the target task for the fine tuning to be effective.

One common approach to improve models' performance with limited data is data augmentation (Shorten & Khoshgoftaar, 2019). This technique is commonly used in other fields of AI, such as computer vision. Attempts have been made to apply data augmentation techniques to textual data (Chen, Tam, Raffel, Bansal, & Yang, 2021), but text 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 (Wu et al., 2022).

This paper is an extended version of Cochran, Cohn, Hastings, and Rouet (2023). The extensions investigate the use of different Large Language Models (LLMs) with different "temperatures" to generate texts to augment the original data, and we compare the results to baseline measurements. Temperature is a parameter that a generative LLM uses to control how much randomness is used when generating text. Higher temperature settings allow for more varied responses. We also compared performance with four different base classifier models to determine if augmentation provides benefits for all of them, or for some more than for others.

As a baseline, we used a "self-augmentation" method where the original data set was replicated to increase training data. This self-augmentation method has been successful in previous research (Cochran, Cohn, & Hastings, 2023), and similar methods have been applied to computer vision with improved model performance (Seo, Jung, & Lee, 2021). 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' textual responses to specific questions, they tend to be on the order of at most few hundred examples. The amount of data obtained was a function of the nature of the texts and the effort required to label the data. 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 (Yogatama et al., 2019). Although some of these models can be good at zero-shot or few-shot learning (Xia et al., 2020), especially when the evaluated texts are relatively "standard", 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 (Gururangan et al., 2020).

These educational data sets are 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 (Schwartz & Stanovsky, 2022). Data augmentation has improved model performance in image processing (Shorten & Khoshgoftaar, 2019). However, that process does not translate directly to text-based models. Studies have used text generation to improve classifier performance by augmenting data to create additional training data artificially (Quteineh, Samothrakis, & Sutcliffe, 2020; Shorten, Khoshgoftaar, & Furht, 2021). The intent was to address the imbalance in data sets and allow smaller data sets to acquire larger data volumes to aid model training. Simple data replication can be used and is referred

to as self-augmentation (Cochran, Cohn, & Hastings, 2023). Looking at techniques beyond self-augmentation, Bayer, Kaufhold, and Reuter (2021) described a taxonomy and grouping for data augmentation types which used replication of the existing data with modifications to the data at the character, word , phrase, and document levels. Cochran, Cohn, Hutchins, Biswas, and Hastings (2022) showed that augmentation using masking, noise, and synonyms can improve classification performance.

The current study continues this research by exploring augmentation using generative AI methods. Several survey papers on text augmentation break down the various types of data augmentation currently being researched (Bayer et al., 2022; Feng et al., 2021; Liu, Wang, Xiang, & Meng, 2020). In the generative method of text augmentation, artificial student responses are generated using a predictive model that infers the response given a text prompt as input. Piedboeuf and Langlais (2024) showed that data augmentation can improve performance for smaller data sets by *generating* text data similar to external data, resulting in fine-tuned models that are higher performing than those with self-augmentation. External data refers to data not generated from paraphrasing or modifying a sentence from the data set, but from data that appears to be provided from an outside source.

The OpenAI API performs NLP tasks such as classification or natural language generation given an input prompt. One of their models is the Generative Pretrained Transformer 3.5 (GPT-3.5) (Brown et al., 2020). For this experiment, we used model "text-curie-001" with 6.7 billion parameters. (Wikipedia, n.d.) A recent study has shown improvement for short text classification with augmented data from GPT, stating that it can be used with additional fine-tuning to improve classification performance (Balkus & Yan, 2023). Additionally, a review by Bayer et al. (2021) noted that GPT was the leading augmentation method among recent papers and may even be able to replicate some instances whose labels were left out of the data set (zero-shot learning).

According to Kumar, Sharma, and Bedi (2024), the most optimal model for NLP tasks, based on size, performance, and resources required for fine-tuning, is Llama from Meta AI

(Touvron et al., 2023). The latest version as of this writing is Llama3 70b. The 70b indicates that the model contains 70 billion parameters. There is a smaller model with 8 billion parameters; however, for this paper, we used the larger 70b version to see if we could get more diverse answers with higher temperatures (Meta, n.d.). In the current study, both GPT and Llama were used for generating augmented responses based on the existing student data and will be compared by analyzing the performance of models fine-tuned by the original student data with the addition of the generated data. The intent was to determine if the language model used matters when generating augmented data, and how much extra data i needed to achieve acceptable performance.

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 (Devlin et al., 2018) and GPT (Brown et al., 2020), are now the industry standard for modeling many NLP tasks. Previous research by Cochran et al. (2022) shows that *BERT-based* transformers work well for text classification of student responses to STEM questions. Therefore, we are using the artificially augmented data sets to fine-tune four types of *BERT-based* models for text classification. Since we have two data sets, one in English and one in French, we used three *BERT-based* multilingual models and one French model as the classifiers of choice.

In this paper, we compared the benefits of generative textual data augmentation from two LLMs for evaluating student textual responses. This evaluation of student responses could be used by teachers and/or given directly to students, but the efficacy of such feedback is beyond the scope of this paper. Accordingly, we evaluate the following research questions and hypotheses.

RQ 1: Can classification performance be improved by augmenting training data with generated responses? Our hypothesis **H1** was that additional generated data would improve model performance for smaller data sets. Determining how large a data set needs to be before it would no longer require data augmentation was out of the scope of this study. Here, we determined if augmentation would work for relatively small data sets.

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

RQ 3: Does altering the response diversity settings of the LLM used to generate student responses affect model performance? Generative models have mechanisms to allow for variability in response generation. Recall that the temperature parameter for GPT and Llama 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 allows the model to choose more "risky" choices from a wider statistical field. **H3** proposes that augmenting the data with slightly more complex answers will generally perform best because generating more complex texts provides additional responses that are not simple paraphrases of the provided student data, allowing the model to generalize.

RQ 4: Does performance ultimately degrade when the model reaches a sufficient level of augmentation? It can be assumed that any augmentation would encounter a plateau such that model performance begins to level off or degrade with additional augmentation (Cochran, Cohn, & Hastings, 2023). When performance levels off, it signals that additional augmentation is not improving performance and that any additional augmentation is providing diminishing returns. When performance degrades, it indicates additional augmentation is either not varied enough, causing the classifier to behave more like self-augmentation, or too varied, causing the knowledge represented in the model to be less focused on the target concept. A counterargument is that generated text that is too complex will be less representative of the student's texts and will "water down" the representation in the model. **H4** was that the performance would level off or degrade with additional augmentation after reaching a peak. We assumed there was a point where additional LLM-generated augmentation from a small data set would add little or possibly degrade performance. **H5** was that the performance would degrade more slowly with higher temperature augmented data sets and thus support the

idea that more risk involved in generated responses was better for more significant amounts of augmentation.

3 Methods

3.1 Cross-Entropy

Entropy measures the disorder or randomness in the data set's label distribution. Looking at the balance between labels in each data set, the entropy can be calculated to indicate the degree of imbalance in a data set.

$$Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

In this formula, n was the number of classes in the data set, and p_i was the probability of class i. It represents how surprising it was to find a particular label in the data set on average. Cross-entropy can be used to compute how distant a distribution was from the balanced distribution.

$$Cross-Entropy(H(p,q)) = -\sum_{x} p(x) \log(q(x))$$

In this formula, p was the balanced probability distribution, and q was the current distribution. Using this calculation, a perfectly balanced data set with two labels would have an entropy of 1.0. When the number of labels increases to four, the entropy goal for a perfectly balanced data set was 2.0. Numbers further away from the ideal value and toward zero indicate a higher degree of imbalance. The cross-entropy has been calculated for each data set to indicate the probability of finding a particular label, indicating the degree of imbalance.

3.2 Data Sets

Two data sets were used in this study. The first data set was 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. Responses were given in French. The participants were asked to read the article and write a passage either to a friend in the "personal" condition or a colleague in the "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 cross-entropy measure for the data for the four possible outcomes. A data set balanced across four labels would have an entropy value near 2.

Table 1 French Student Response Data Split for the Opinion Concept

	No Opinion	No Link	Partial Link	Link Exists	Cross-Entropy
Count	**118**	13	25	7	0.619

The second data set was obtained from a study on students learning about rainwater runoff with responses from 95 6th-grade students in the southeastern United States (Basu et al., 2022; Hutchins et al., 2021; McElhaney et al., 2020; Zhang et al., 2020). Responses were given in the English language. In this study, three questions related to a fictitious student named "Libby" were asked. The six concepts came from the following three questions where their associated concepts were:

Question 1: What do you think the different-sized arrows in Libby's model could mean? Question 1 had one correct response: the size of the arrows indicates the amount of water.

Question 2: What are two things that you would change about Libby's model to explain where the water goes? This question focused on finding errors in the model, explaining the error, and providing the correct answer. Good answers to the question would included two concepts: the size of the runoff and absorption arrows should sum to the size of the rainfall arrow (conservation of matter), and the direction of the runoff arrow should be pointing downhill.

Question 3: What are two things that Libby's model does a good job of explaining? The ideal answer to this question could address three concepts: rainfall either gets absorbed or becomes runoff, the sizes of the arrows in the diagram correspond to the different amounts of water, and rainfall is the origin of the water.

The six concepts from the Rainwater data set are shown in Table 2. They were each modeled individually as a binary classification task. Student responses that included the corresponding concept were coded as **Present**. Responses were otherwise coded as **Absent**.

Table 2 Concepts present in each question for Rainwater Runoff

Question	Concept	Description
1	C1	Arrow size indicates the amount of water
2	C2a	Size of runoff and absorption arrows should sum to the size of rainfall arrow
2	C2b	Direction of runoff arrow should be pointing downhill
3	C3a	Model demonstrates rainfall either absorbed or becomes runoff
3	C3b	Model illustrates where water is coming from
3	C3c	Model uses arrow size to indicate water amount

As previously mentioned, many small educational data sets are imbalanced. Table 3 shows the label quantities indicating the scarcity of data and the corresponding cross-entropy, showing the degree of imbalance in the Rainwater concepts. A data set balanced across two labels would have a cross-entropy of 1.

Table 3 Rainwater Runoff Student Response Data Split per Question

Concept	Absent	Present	Cross-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

3.3 Augmentation Approach

Tables 1 and 3 show the label quantities for each concept (with the majority label in bold), along with the cross-entropy. Cochran et al. (2022) showed that balancing the data set was

imperative to get a reliable performance result when fine-tuning with small educational data sets. Each label would have equal quantities to balance a data set, and the cross-entropy values would be at or near 1.0 for binary labels and closer to 2.0 for a data set with four labels.

In this work, we define an augmentation level of *0x* to indicate when all labels have the same quantity as the majority quantity of reference for that data set. Therefore, the data set was balanced. That is, 0x does not mean that there was no augmentation, but that augmentation was used to add enough data per label to equal the majority quantity in the original set for all labels. For example, with the French data with a majority quantity of 118 for "No Opinion", we needed 105 additional examples of "No Link", 93 of "Partial Link", and 111 of "Link Exists" to reach 0x, or 118 examples per label. Additional augmentation was then applied in multiples of the majority quantity, starting at 1x and going up to 100x, or 100 times the majority quantity for that data set.

With GPT, we generated data using the prompt "paraphrase this sentence" and inserted each actual student response to fill in the rest of the language prompt. We repeated this for each student's response. For Llama, a similar approach was used, except that the instructions for generating text were placed in the "system content" section and the example prompt was placed in the "prompt" section of the API. The Llama system content used for English samples was "Your task is to generate a response similar to the text provided, without saying the same exact text." For the French samples, the system content was "You will be provided with a sentence in French, and your task is to generate a response similar to the text provided without saying the same exact text, and return it in French."

The data was generated, stored, and used directly to fine-tune 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 properly.

3.3.1 Temperature

Both GPT and Llama provide a method for varying the degree of freedom in generating text by adjusting the input parameter, "temperature". *Temperature sampling* balances predictability and creativity during text generation. Higher temperature settings allow less likely tokens to be selected during text generation, whereas a lower temperature value increases the confidence in statistically most likely choices. GPT allows a floating-point temperature range of 0.0 to 1.0, whereas Llama has a temperature range of 0.0 to 2.0. In this study, we performed tests at temperature values of 0.1, 0.5, and 0.9 for GPT and 1.5 and 2.0 for Llama. The results were then analyzed to determine if temperature was an important factor in text generation in that it affected fine-tuned model performance.

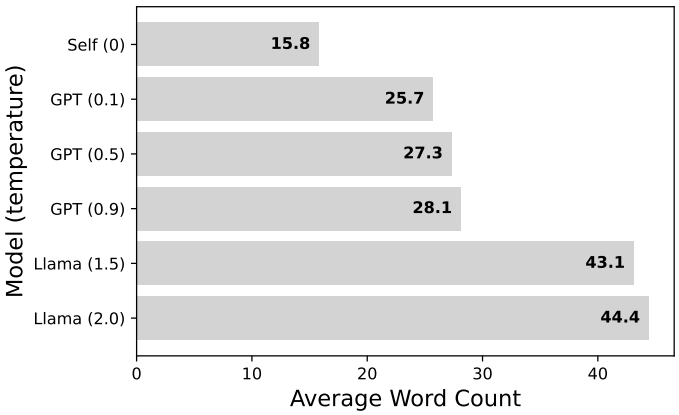

Fig. 1 Average Word Count of Generated Responses based on Model and Temperature Setting.

The different temperature settings for the models had a significant effect on the length and complexity of the generated sentences. Figure 1 shows the average word count for each text generation method and temperature. The "Self (0)" bar shows the average length of the original student texts. Examples of generated texts are provided in the following section.

3.3.2 Sentence Complexity

As mentioned above, sentences generated by GPT and Llama showed differing profiles. The GPT-generated responses were similar in length and complexity to the student's responses. However, the Llama-generated responses were much longer and more complex than those generated by GPT. The Llama model with a temperature = 2.0 showed the highest complexity and longest length of all models and temperatures that generated responses. In this section, we describe how we measured sentence complexity.

There are many different metrics for readability. For this study, we used the Gunning fog index,[1] developed in 1952. Although the Gunning fog metric was developed for use on English texts, its definition is simple enough to apply to other languages with the same alphabet. It also allows us to directly compare (albeit approximately) the readability levels between languages. For this research, we are less concerned about the absolute readability levels for the texts and more interested in the relative complexity levels of the generated texts.

The Gunning fog index is meant to reflect the number of years of schooling that a student would need to be able to read the text. It is defined in terms of the number of *words* in the text, the number of *sentences*, and the number of *complex words*, i.e., words with 3 or more syllables (not counting common suffixes). It is calculated as:

$$0.4\left[\frac{words}{sentences}+100\frac{complexWords}{words}\right]$$

With respect to the United States education system, fog index values from 1 – 6 would be associated with elementary school (i.e., up to roughly age 12), scores from 7 – 12 denote middle to high school, 13–16 are college level, and 17 or above would indicate graduate or post-graduate level.

In addition to the limitation mentioned above that the Gunning fog formula was developed for English-language texts, other limitations are that its accuracy is best with texts of over 100 words. It also relies on some simplifying assumptions. For example, some words of three or

[1] https://en.wikipedia.org/wiki/Gunning_fog_index

more syllables words are easily understood by younger students. But again, we are primarily interested in comparing the complexity of texts rather than for computing the most accurate absolute readability values.

Table 4 shows an original student-generated correct answer to rainfall question 3a and LLM-generated text responses with different temperatures. It also gives the Gunning fog readability score for each. Table 5 shows examples from the French Video Games data set, also with the Gunning fog readability scores.

Table 4 Sample of Generated Student Responses for Rainwater Runoff (Concept 3a). (Note: GF is the Gunning-Fog Readability score for the sample.)

Generator	Temp	Sample for concept 3a	GF Score
Original		I think it means the amount of rainfall or water that is either falling, being soaked in, or running off.	8.0
GPT	0.1	The word "rain" means the amount of rainfall or water that is either falling, being soaked in, or running off.	8.8
GPT	0.5	The word "rain" usually refers to precipitation, which is water that falls from the sky, is soaked in, or runs off.	12.7
GPT	0.9	I think what you are trying to say is that rain means liquid water that falls from the sky, gets soaked up by the ground, or runs off.	11.2
Llama	1.5	I believe it refers to the sum of precipitation, absorption, and drainage of water in a specific area, encompassing water that's falling from the sky, being absorbed into the ground, and flowing over the surface.	19.7
Llama	2.0	So you're referring to the total sum of water fluxes involved in the hydrologic cycle, encompassing precipitation, infiltration, runoff, and perhaps also evapotranspiration, which together govern the flow of water within watersheds and ecosystems.	25.4

To determine if the two generators with their different temperature scores were creating texts with significantly different complexity values, we compared the readability scores of the texts using t-tests. Our initial observations suggested two things:

- Higher temperature settings resulted in more complex sentences (i.e., higher readability indices).
- Llama3 70b generated more complex sentences than GPT 3.5 did.

Accordingly, we performed t-tests to evaluate these observations. More specifically, for each separate question in our French and English data sets and for each generator / temperature combination, we performed a two-sample, one-sided t-test to test the hypothesis that

Table 5 Sample of Generated Student Responses for the French Study by Model and Temperature (T). (Note: GF is the Gunning-Fog Readability score for the sample.)

Generator	T	Sample	GF score
Original Text		Je penses que d'apres les sources telles que la tribune et le monde avec des resultats d'experiences assez divergeants, la violence ne serait pas forcement liée aux jeux vidéos mais plutot peut etre à la nature meme de la personne qui joue. peut etre elle ne fait que faire ressortir ce qu'est réellement la personne en question. c'est donc quelque chose qui depend de plusieurs variables.	13.6
GPT	0.1	Je pense que les résultats des expériences telles que la parole et le monde montrent que la violence n'est pas forcément liée à des jeux vidéo, mais plutôt à la nature du joueur en les jouant. Peut-être révèlent-ils ce qui se cache sous la surface de l'individu. Cela dépend des facteurs.	8.4
GPT	0.5	Je pense que les résultats d'expériences indiquent que la violence n'est pas forcément liée aux jeux vidéos, mais plutôt à la nature même de la personne qui les joue. Peut-être qu'elle ne fait que révéler ce qui est réellement la personne, et cela dépend de plusieurs variables.	13.7
GPT	0.9	Je crois que selon les sources telles que la tribune et le monde, avec des résultats d'expériences assez différents, la violence ne serait pas forcément liée aux jeux vidéo, mais plutôt peut-être à la nature même de la personne qui joue. Peut-être qu'elle ne fait que ressortir ce qu'est réellement la personne en question. C'est donc quelque chose qui dépend de plusieurs variables.	11.6
Llama	1.5	Je crois que, nhu ce que suggèrent des études de différentes sources, notamment La Tribune et Le Monde, les résultats sont contradicoires et qu'il est difficile de drawers une corrélation directe entre les jeux vidéos et la violence. Il est possible que ces derniers ne fassent que révéler certaines tendances préexistantes chez les joueurs, plutôt que d'en être la cause directe. De plus, il est évident que d'autres facteurs interviennent, ce qui rend la compréhension de cette problématique complexe.	15.1
Llama	2.0	Il est intéressant de noter que les recherches et les avis d'experts tels que ceux de Le Monde diplomatique et de nombreuses enquêtes montrent des résultats contradictoires, ce qui suggère que la violence n'est pas une consequence directe des jeux vidéos, mais plutôt qu'elle constitue une réponse à différents stimulus, notamment lié à la personnalité très especifique du joueur en lui-même. D'autres éléments, tels que l'environnement familial, vient en complément porter cette compréhension plus globale signalant ainsi ne plus ne qu'une possibilité d'eremble mais diversite possibilités.	26.0

the generated texts' readability scores were the same as those of the "simpler" generators, i.e., those with lower temperature settings. We used a one-sided t-test because we expected that the simpler generators would generate simpler sentences. We used 1000 texts randomly sampled from those returned by each generator / temperature combination.

Table 6 shows the significance values for the French video games question and for English rainwater questions 2a and 2b. The generator / temperature combinations are shown in the first row and first column for each subtable. The second row and column show the means and

standard deviations for the readability scores. Each cell within the table shows the significance level, with "***" indicating $p < 0.001$, and "*" indicating $p < 0.05$.

For example, the French question portion at the top of Table 6 shows that, in every case, the mean readability of texts generated by a more "complex" generator (i.e., one with a higher temperature) were significantly higher than those from a "simpler" generator. All the differences were highly significant ($p < .0001$) with the exception that texts from GPT 3.5 with a temperature of 0.9 had a mean readability of 13.9 ($SD = 3.8$), which was only significantly higher at a $p < .05$ level from those generated by GPT 3.5, Temp=0.5, $M = 13.6$, $SD = 3.9$.

Table 6 Significance of Gunning fog readability score differences between Generators (with given temperatures), for French videogames question and Rainwater Questions 2a and 2b. Each row indicates the probability that the distribution of readability scores of 1000 randomly selected sentences is not greater than those generated by the Generator above. (*** indicates $p < 0.001$, * indicates $p < 0.05$)

		French Games Question			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	12.8 (4.8)	13.6 (3.9)	13.9 (3.8)	16.8 (3.5)
GPT 0.5	13.6 (3.9)	***			
GPT 0.9	13.9 (3.8)	***	*		
Llama 1.5	16.8 (3.5)	***	***	***	
Llama 2	23.1 (9.5)	***	***	***	***
		Rainwater Question 2a			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	8.8 (3.4)	8.9 (3.4)	9.4 (3.5)	16.1 (4.8)
GPT 0.5	8.9 (3.4)	0.24			
GPT 0.9	9.4 (3.5)	***	***		
Llama 1.5	16.1 (4.8)	***	***	***	
Llama 2	17.2 (5.2)	***	***	***	***
		Rainwater Question 2b			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	9.0 (3.3)	9.2 (3.3)	9.8 (3.5)	16.6 (4.5)
GPT 0.5	9.2 (3.3)	0.10			
GPT 0.9	9.8 (3.5)	***	***		
Llama 1.5	16.6 (4.5)	***	***	***	
Llama 2	17.9 (8.3)	***	***	***	***

Table 7 shows the significance levels for readability between sentences for English-language rainwater questions 1 and 3a, 3b, and 3c. Here, we can see that there were no

Table 7 Significance of Gunning fog readability score differences between generators, for Rainwater Questions 1 and 3a – 3c.

		Rainwater Question 1			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	8.2 (3.1)	8.3 (3.4)	9.0 (3.7)	14.1 (5.1)
GPT 0.5	8.3 (3.4)	0.10			
GPT 0.9	9.0 (3.7)	***	***		
Llama 1.5	14.1 (5.1)	***	***	***	
Llama 2	15.6 (6.2)	***	***	***	***
		Rainwater Question 3a			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	10.0 (3.7)	9.6 (3.6)	10.2 (3.5)	15.9 (4.9)
GPT 0.5	9.6 (3.6)	0.98			
GPT 0.9	10.2 (3.5)	0.10	***		
Llama 1.5	15.9 (4.9)	***	***	***	
Llama 2	16.9 (5.4)	***	***	***	***
		Rainwater Question 3b			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	10.2 (3.4)	10.1 (3.7)	10.1 (3.4)	15.8 (5.0)
GPT 0.5	10.1 (3.7)	0.78			
GPT 0.9	10.1 (3.4)	0.60	0.30		
Llama 1.5	15.8 (5.0)	***	***	***	
Llama 2	16.8 (5.7)	***	***	***	***
		Rainwater Question 3c			
Gen/Temp		GPT 0.1	GPT 0.5	GPT 0.9	Llama 1.5
	Mean(SD)	10.1 (3.5)	9.7 (3.6)	10.0 (3.5)	15.8 (5.0)
GPT 0.5	9.7 (3.6)	0.99			
GPT 0.9	10.0 (3.5)	0.83	*		
Llama 1.5	15.8 (5.0)	***	***	***	
Llama 2	16.8 (5.2)	***	***	***	***

significant differences between the readability of texts generated by GPT 3.5 at different temperature settings for several questions. Llama did, however, generate significantly more complex texts.

After we created artificial student responses like these using the two LLMs with five different temperature settings, we used those artificial responses to augment the original (small) data sets and balance the data between the outcome labels. Then, the augmented data sets were used to fine-tune classifiers and test model classification performance to see if the performance improved.

3.4 Model Selection

Since we had data sets in two different languages, we chose three multilingual models and one French model to compare the effect of input language when performing fine-tuning. The chosen models and their original intended tasks are shown in Table 8. We downloaded pre-trained models from HuggingFace because multilingual models are available, including French and English languages, perform text classification as a downstream task, and have a high number of downloads. We chose BERT Base multilingual uncased from this narrowed list as a general model. Next, we chose the Microsoft *Multilingual L12 H384* model due to its performance gains over the base BERT model and its improved ability for fine-tuning (Wang et al., 2020). Additionally, we selected *jeveuxaider/activity-classifier* (HuggingFace, n.d.), a French multi-class classifier that has a high number of downloads, and *classla/xlm-roberta-base-multilingual-text-genre-classifier* (Kuzman, Mozetič, & Ljubešić, 2023) which performs text classification based on xlm-roberta-base and fine-tuned on a combination of three genre data sets.

Table 8 Models Chosen for Evaluation

Model	Intended Target Task
nlptown: BERT Base multilingual uncased	Sentiment Analysis
Microsoft: Multilingual L12 H384	Language Understanding and Text Generation
jeveuxaider: activity-classifier	French Text Classification
classla: xlm-roberta-base-multilingual-text-genre-classifier	Multilingual Text Classification

We fine-tuned each pre-trained model using original data augmented with generated data created by the LLMs. During model fine-tuning, we used the following hyperparameters: max_len 128, epochs 3, batch size 32, optimizer Adam, learning rate 1 e-5, warmup 0.1, cost sensitivity 0. Twenty percent of the data was held out from the original data set for testing purposes using a different random seed for each experiment. No augmented data were included in the test sets that we used to measure model performance. Any augmented data generated from the test data was excluded from the training set for that particular model so it would not inadvertently provide context around the withheld test data set, creating an inflated

performance. Six augmentation methods (self + five LLM / temperature combinations) were applied to each of the four pre-trained BERT models, which were fine-tuned for each of the six rainwater concepts and one French concept. Ten seeds were used for each model, and nine augmentation levels were chosen (0, 1, 3, 8, 21, 34, 55, 89, and 100). This 6 x 4 x 7 x 10 x 9 combination resulted in 15,120 (!) separate *BERT-based* models that were fine-tuned and evaluated for this study. We used the micro-F_1 metric as the performance measurement. Each model's performance was averaged over the ten seeds.

3.5 Data Creation, Model Fine-Tuning and Testing

Building models for this paper was computationally intensive, and online resources for model building, such as Google Colab and IBM Watson, were inadequate to perform this task. For this reason, a machine was created to perform all calculations for this set of experiments consisting of an AMD 1900 Threadripper CPU and a single NVIDIA GeForce RTX 2080 Ti graphics card with 11 GB of memory. This machine fine-tuned all 15,120 models individually with varying levels and types of augmentation from the two data sources. Generating the required amount of text for augmentation from the generative LLMs and fine-tuning each model took over eight weeks of continuous, 24/7 GPU usage.

3.6 Baseline Evaluation

We evaluated two different baseline models for each concept. For the *a priori* model, we chose the majority classification label for each concept. In other words, we simply chose the majority label as the guess for the classification. For our *unaugmented* baseline, we applied BERT prototypically without data augmentation or balancing. The baseline performance results and the BERT results on augmented data are included in the tables in the next section.

4 Results

Tables 9, 10, 11, and 12 present summaries of the results for each model. Each row corresponds to a concept: one for the French data set and six for the English 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 micro-F_1 scores for each concept using self-augmented or LLM-generated data. The highest performance level for each concept is shown in **bold**. The two rightmost columns indicate the augmentation level and temperature that were used to achieve maximum performance. Recall that the temperatures for GPT are on a 0 to 1 scale, whereas the Llama temperatures are on a 0 to 2 scale.

Table 9 **Microsoft *Multilingual L12 H384*** Model Performance (micro-F_1) of Baseline vs All Augmented Models

	% Maj.	Baseline		Max Performance				
Concept	Label	*a priori*	Unaug.	Self	GPT-3.5	Llama3	Aug.	Temp
French	73	0.730	0.371	0.651	0.612	**0.894**	100x	1.5
C1	89	**0.890**	0.735	0.789	0.816	0.853	-	-
C2a	73	0.730	0.757	0.932	0.816	**1.000**	21x	2
C2b	67	0.670	0.547	0.636	0.884	**0.936**	55x	2
C3a	54	0.540	0.532	0.721	0.832	**0.879**	100x	2
C3b	77	0.770	0.684	0.926	**0.947**	0.942	21x	0.9
C3c	60	0.600	0.568	0.742	0.832	**0.879**	100x	2

Table 10 **Jeveuxaider *Activity-Classifier*** Model Performance (micro-F_1) of Baseline vs All Augmented Models

	% Maj.	Baseline		Max Performance				
Concept	Label	*a priori*	Unaug.	Self	GPT-3.5	Llama3	Aug.	Temp
French	73	0.730	0.575	0.667	0.789	**0.924**	100x	1.5
C1	89	**0.890**	0.789	0.753	0.789	0.873	-	-
C2a	73	0.852	0.842	0.895	0.905	**0.973**	21x	2
C2b	67	0.670	0.737	0.789	0.858	**0.921**	55x	2
C3a	54	0.540	0.460	0.763	**0.815**	0.805	55x	0.1
C3b	77	0.770	0.947	0.868	**0.974**	0.952	21x	0.9
C3c	60	0.600	0.527	0.747	0.868	**0.905**	100x	2

Table 11 **Classla *xlm-roberta-base-multilingual-text-genre-classifier*** Model Performance (micro-F_1) of Baseline vs All Augmented Models

	% Maj.	Baseline		Max Performance				
Concept	Label	*a priori*	Unaug.	Self	GPT-3.5	Llama3	Aug.	Temp
French	73	0.730	0.371	0.681	0.821	**0.939**	100x	2
C1	89	0.890	0.789	0.795	0.816	**0.895**	89x	2
C2a	73	0.852	0.842	0.895	0.916	**1.000**	8x	2
C2b	67	0.670	0.737	0.789	0.889	**0.947**	34x	2
C3a	54	0.540	0.460	0.726	0.816	**0.867**	89x	2
C3b	77	0.770	0.947	0.947	0.968	**0.974**	100x	2
C3c	60	0.600	0.525	0.721	0.842	**0.894**	100x	2

Table 12 **Nlptown *BERT Base multilingual uncased*** Model Performance (micro-F_1) of Baseline vs All Augmented Models

	% Maj.	Baseline		Max Performance				
Concept	Label	*a priori*	Unaug.	Self	GPT-3.5	Llama3	Aug.	Temp
French	73	0.730	0.371	0.648	0.821	**0.873**	100x	1.5
C1	89	**0.890**	0.789	0.789	0.821	0.863	-	-
C2a	73	0.852	0.842	0.842	0.905	**0.968**	21x	2
C2b	67	0.670	0.737	0.753	0.889	**0.937**	89x	1.5
C3a	54	0.540	0.544	0.747	0.732	**0.842**	89x	2
C3b	77	0.770	0.947	0.947	0.847	**0.953**	100x	2
C3c	60	0.600	0.525	0.753	0.905	**0.926**	100x	1.5

Figure 2 illustrates how each of the six augmentation methods (self + five temperature/LLM combinations) affected model performance as more augmentation was used to fine-tune each of the four types of models using the French data. 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 API to provide varied responses, as previously discussed. For Llama, the temperature settings were 1.5 and 2. At levels of augmentation less than 20x, classification performance with all of the BERT models was generally unreliable.

Figures 3, 4, 5, and 6 show how each model's performances varied with training data using the rainwater data in English, showing different augmentation types of self, and the five LLM / temperature combinations used to generate text. Here, too, augmentation levels below 20x resulted in unreliable classifications in the majority of cases. The classifications were quite good, but the performance depended significantly on the particular concept being classified.

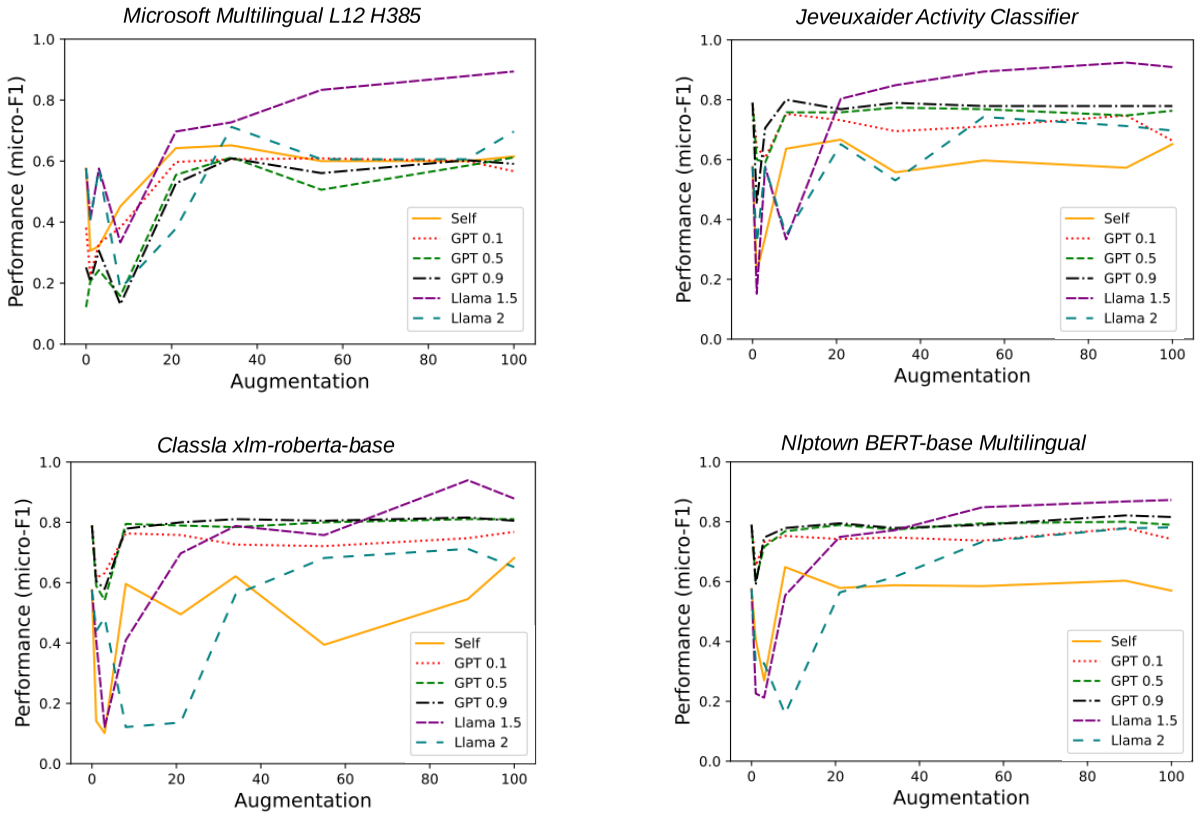

Fig. 2 Classification performance (micro-F_1) with augmentation for the four BERT models. (Note: The x-axis shows the level of augmentation applied from 0x to 100x.)

Although there was more variability here, Llama-generated text with the highest temperature = 2 generally resulted in the best performance.

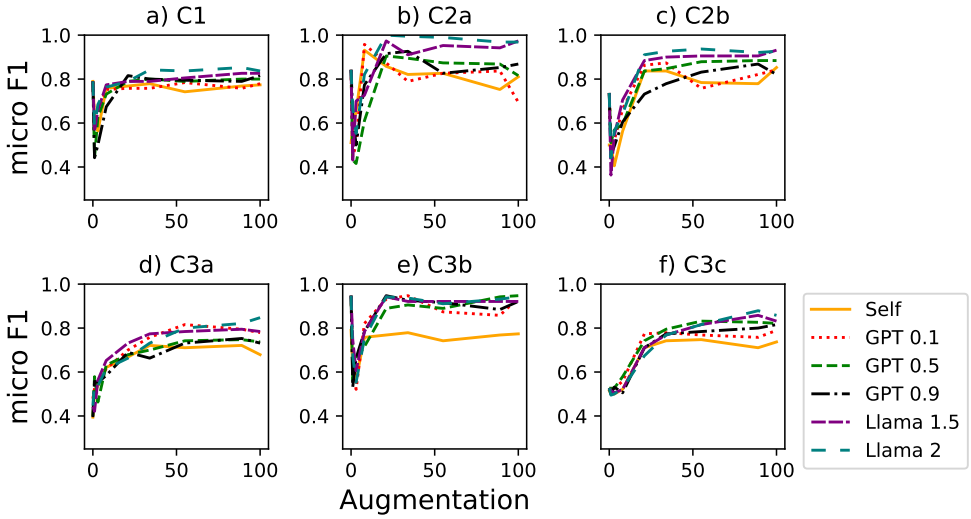

Fig. 3 Rainwater Runoff Model Performance (micro-F_1) per Augmentation Amount for each Augmentation Type for the **Microsoft *Multilingual L12 H384*** Model. (Note: The *x*-axis shows the level of augmentation applied from 0x to 100x.)

Table 13 shows the peak performance for each concept, along with the corresponding augmentation details. As mentioned above, for the French data, Llama with a temperature setting of 1.5 provided the best performance, at a high level of augmentation, 89x. The Classla *xlm-roberta-base-multilingual-text-genre-classifier* model produced the highest classification accuracy, but the other models also performed well, as shown in Figure 2.

For the English concepts, again the Llama LLM generated the most useful sentences for augmenting the training set, with the exception of concept C3b, where the GPT-generated sentences performed equally well, with a very high micro-F_1 of 0.974. For the English sentences, however, a higher temperature (2 for Llama, .9 for GPT) produced more useful texts. High augmentation levels were also beneficial, with one notable exception; both the

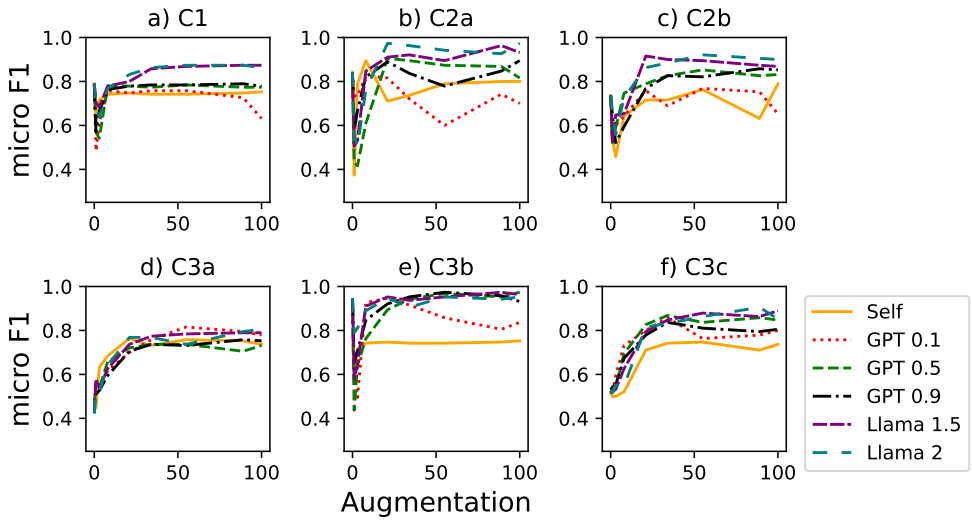

Fig. 4 Rainwater Runoff Model Performance (micro-F_1) per Augmentation Amount for each Augmentation Type for the **Jeveuxaider *Activity-Classifier*** Model. (Note: The x-axis shows the level of augmentation applied from 0x to 100x.)

Microsoft *Multilingual L12 H384* model at 21x augmentation and the Classla *xlm-roberta-base-multilingual-text-genre-classifier* at 8x augmentation were able to correctly classify *every item in the test* for each of the 10 random seeds that were used.

Table 13 Best Performance (micro-F_1) Achieved for Each Model Chosen for Evaluation

Concept	Model	Aug LLM	Aug Level	Temperature	Max Performance
French	Classla	Llama	89x	1.5	0.939
C1	Classla	Llama	89x	2	0.895
C2a	Microsoft and Classla	Llama	21x and 8x	2	1.000
C2b	Microsoft	Llama	55x	2	0.936
C3a	Microsoft	Llama	100x	2	0.879
C3b	Microsoft and Classla	GPT and Llama	21x and 100x	0.9 and 2	0.974
C3c	Nlptown	Llama	100x	1.5	0.926

5 Discussion

Recall **RQ 1**, which asked if LLM-augmented data sets would improve classifier model performance over unaugmented data sets. Our hypothesis **H1** stated that additional augmented data would improve model performance for smaller data sets, compared to unaugmented

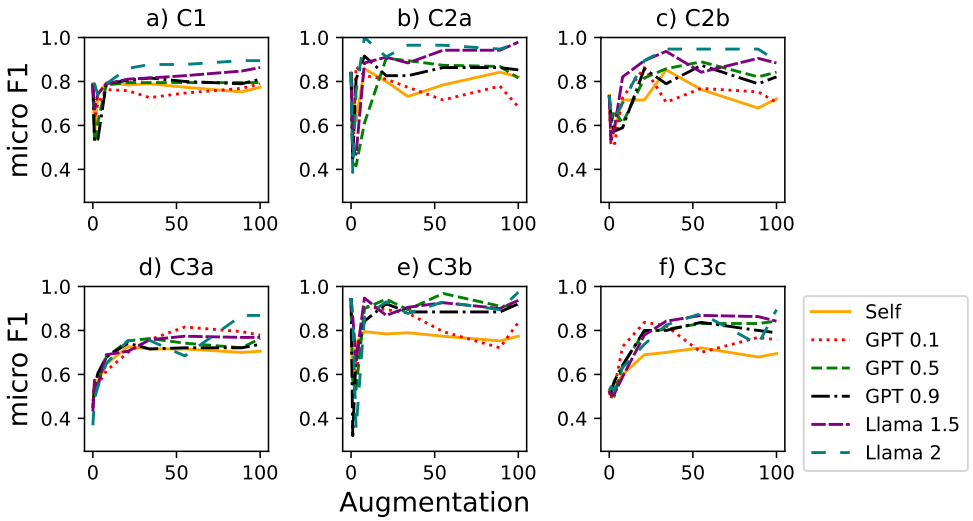

Fig. 5 Rainwater Runoff Model Performance (micro-F_1) per Augmentation Amount for each Augmentation Type for the **Classla *xlm-roberta-base-multilingual-text-genre-classifier*** Model. (Note: The *x*-axis shows the level of augmentation applied from 0x to 100x.)

training sets. The current study showed that models trained on LLM-augmented data outperformed the unaugmented models for all seven concepts, no matter which of the four classifiers was used.

However, the *a priori* computation, which always selected the majority label, outperformed augmented models on one of the data sets, C1. This is likely due to the severe imbalance of the data set prior to augmentation. The C1 data set had 89% of the data marked with the majority label and an entropy value of 0.485. Note in Table 2 that the first concept was simple, and most students would likely get this question correct as it simply asks what the size of the arrow on a diagram indicates (i.e., the amount of water). Whenever there are simple or, on the contrary, difficult questions, there will likely be an imbalance in the data set. In our testing, we observed that the more straightforward the question, the more imbalanced the data was because the majority of students got it correct. Incorrect responses were few, making it difficult to generate enough unique samples to expand the minority class, resulting in a minority class augmentation pool that resembles self-augmentation, which has been shown

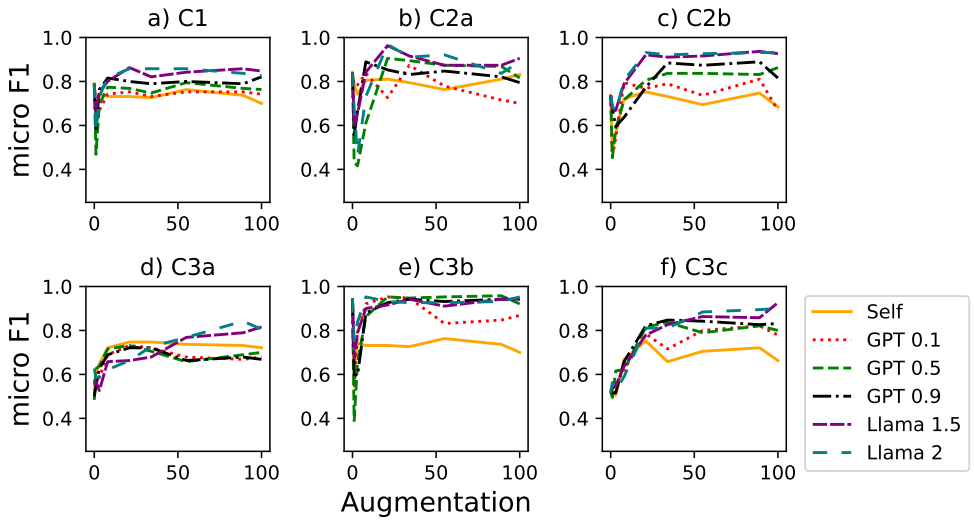

Fig. 6 Rainwater Runoff Model Performance (micro-F_1) per Augmentation Amount for each Augmentation Type for the **Nlptown *BERT Base multilingual uncased*** Model. (Note: The *x*-axis shows the level of augmentation applied from 0x to 100x.)

to only improve performance slightly more than no augmentation. The resulting classifier suffers in performance due to the minority class deficiency.

In our testing, only one combination of classifier model and augmentation method increased performance enough to outperform blind guessing of the majority label. We concluded from our experimentation that data sets with an entropy of less than 0.5 were challenging to augment sufficiently (i.e., with enough generated data) to create a classifier that would outperform simply guessing the majority label. This shows that the initial classifier model selection was important because models do not rise in performance equally, and experimentation is necessary to determine which classifier models will perform the best. Because classifier model performance is difficult to predict, we recommend testing to find the best-performing classifier model by augmenting the data and applying additional fine-tuning. It is also essential to augment the training set and continue to train the model with this data until peak model performance is achieved. The results supported the hypothesis for the unaugmented baseline, but concept C1's *a priori* performance outperformed augmented classifier performance in three of the four classifiers trained.

RQ 2 asked if using LLM-generated responses would outperform self-augmentation when training models for classification. Our hypothesis **H2** stated that training with LLM-generated responses would outperform simple replication of existing data. This hypothesis was supported by our results which showed that maximum performance was achieved using generated responses for all seven concepts. LLM-generated responses outperformed self-augmentation because the generated texts were longer and more complicated. Higher performance was achieved above augmentation levels of 20x, most notably by the Llama-augmented texts with a temperature = 2 for binary classification. However, it is interesting to note that since the French data was a multi-class classification (i.e., with four possible labels), performance was better with a lower temperature = 1.5. In all cases, performance by self-augmentation plateaued more quickly and often decreased in performance with additional augmentation, indicating that classifier model training is more stable when using generated text. The self-augmentation method led to unstable models that varied in performance at higher levels of augmentation.

RQ 3 asked if adjusting the randomness of generated student responses would affect model performance. **H3** proposed that augmenting the data with slightly more risky (i.e., varied) answers would provide the best performance in general. Examining the maximum performance at each augmentation level revealed that Llama with temperature = 2 was the top performer among the five generator / temperature combinations used for student response generation. Figure 1 above shows the average word length of the generated responses based on the model and temperature combination. The average student response was 15.8 words on average, indicated by the "Self(0)" value. Tables 4 and 5 above show a sampling of the sentence complexity, and Figure 1 shows the word count, demonstrating that Llama with a temperature = 2 generated the longest and most complex texts. Figures 2 and 3 and Table 9 demonstrated that models using Llama-generated texts with a temperature = 2 also resulted in the highest performance. This setting generated texts that were 44.4 words on average, whereas GPT with a temperature = 0.1 generated responses of 25.7 words on average, or

about 60% as long. The longer set of generated texts from Llama appears to lead to improved performance in almost every case. Tables 4 and 5 above also showed an increase in response complexity, paralleling the increase in word count.

Interestingly, GPT-generated responses did not increase in complexity with a higher temperature value like the Llama-generated responses, as shown in Table 6. The Gunning fog readability score decreased with larger temperature values for GPT, especially for concepts 3a, 3b, and 3c. This could be attributed to the greater complexity of the answers associated with these concepts. Recall Table 2 where the first concepts (1, 2a, 2b) were the questions were simpler, asking for either direction or quantity measures. Question 3 required a more advanced response as the students needed to show an understanding of how the model represented each of the concepts.

It appeared that the responses from GPT more closely resembling a rewording of the original text, as directed by the prompt instructions. Llama, on the other hand, went beyond rewording and attempted to explain the prompt with higher temperature values and, therefore, generated much more complex completions. This indicates that future performance gains may be had with additional prompt engineering, where input to the LLM is modified to guide the generator to producing more suitable alternative texts for augmenting the training set.

The charts presented in Figures 2, 3, 4, 5, and 6 above showed that each data set augmented by the different models and temperatures varied in performance, but the top performing models used augmented data from Llama with a temperature = 2. Toward higher levels of augmentation, the more risky, variable generation continued to increase model performance. These results support the hypothesis that temperature settings do affect model performance when the responses generated have a larger word count and are more complex. From this set of experiments, the generated responses with the highest sentence complexity pushed classification model performance to the highest levels for the English rainfall data

set. However, the lower temperature = 1.5 for Llama produced the highest classifier performance for the French video games data set. For this reason, this hypothesis was only partially supported.

Finally, **RQ 4** asked if performance would ultimately level off or degrade when the training set reached a sufficient level of augmentation. **H4** posited that the performance with additional augmentation would either level off due to diminishing returns, or even degrade due to a training set that became less centered on the target concept. In self and GPT-augmented models tested, performance peaked and degraded after 55x to 89x augmentation. For the Llama-augmented models, performance continued to rise as the augmentation neared 100x, suggesting more augmentation could be applied above 100x to achieve even higher performance. This hypothesis was partially supported as not all models degraded as augmentation amounts reached 100x. They will likely eventually level off, but this evidence does not fully support the hypothesis. Table 13 above showed the augmentation level at which performance peaked for each concept, showing the model and augmentation details.

H5 suggested that the performance would slowly degrade with higher temperatures and more varied generated responses. This was true up to 100x for our experiments. As shown in Figures 2 and 3 above, using texts from the Llama generator with the highest temperature = 2, performance rose with increasing amounts of augmentation, but did not show the decrease that the other combinations of generator and temperature did at higher augmentation levels. Due to this observation, this hypothesis was partially supported by the data. As augmentation increased, the "self" method peaked and began to decline in performance with additional training data added, where the augmented models using GPT did not drop off as much. With Llama-generated texts, the peak in performance was seen at higher amounts of augmentation, and many did not degrade. This indicates the model was more tolerant of generated data, especially when the text was more complex and lengthy than the original text. It was difficult to determine the exact level of augmentation required as a universal constant. From this study, the exact level of augmentation required to gain a sufficient level of performance varied with the data set and model chosen.

6 Conclusion

This study aimed to determine if we could improve classification of student answers by augmenting small, often imbalanced, training sets with texts generated by GPT-3.5 and Llama3 70b at different temperature settings. We used one French and three multilingual *BERT-based* classifier models, trained them using two different data sets in two languages augmented by six different methods, one using self-augmentation and five with GPT and Llama augmentations, and compared the results to two baseline models. Our results showed that including texts generated by these two LLMs produced marked performance improvements over self-augmentation alone. The texts generated from Llama with higher temperature settings resulted in the best overall classifier performance.

Another objective of this study was to determine if temperature settings in response generation would affect classifier performance. Our results showed that, especially for Llama, higher temperature settings produced texts that were longer and more complex. Temperature settings close to zero pushed the LLMs to generate texts much closer to the original text in the prompt. The higher the temperature setting, as it approaches infinity, the closer the response generation comes to universal sampling, meaning the probability is ignored and every possible completion is equally likely to be generated. Both GPT and Llama have limits on the temperature, but higher temperatures allow for completions with a lower probability of being the best completion. Most of the models augmented with Llama-generated responses with temperature = 2 pushed the performance beyond what was achieved by other augmentation methods because the text generated from this combination of model and temperature provided more robust, varied training sets that included longer, higher-complexity texts. This type of augmentation allowed the models to continue to increase performance with more augmented data before performance began to show signs of leveling off. The performance also maintained a stable increase longer than those with lower temperatures, including self-augmentation.

We tested only one classifier model in the original conference paper that we are extending here (Cochran, Cohn, Hastings, & Rouet, 2023). In this paper, we include three additional

models. This was done to see if model selection was essential or if augmentation would overshadow any performance differences between classifiers. As shown in Table 13 above, the Classla model performed well for the severely unbalanced data sets for concept C1 and the French data, whereas the Microsoft model performed higher in most other concepts. Since the Jeveuxaider classifier was trained on French text, it might be assumed that this model would perform best on French data. As it turns out, this model performed well but was not the overall highest performer. Selecting a model was not as easy as picking one trained in the language of the data set and fine-tuning it on the target task. Some models that may not appear to be the ideal fit initially will sometimes perform better than those that fit the problem well.

These empirical tests show that augmentation using LLM text generation can markedly improve performance over unaugmented or self-augmented models. It also showed that generated texts can vary in quality, and the more complex and longer the word length of the text is, the better the performance of classification models in general.

Imbalance in the data set is problematic. Both data sets had areas where they were severely imbalanced. For the English data set, specifically with concept C1, even the highest amount of augmentation with GPT-generated texts was insufficient to outperform the simple *a priori* baseline due to the high level of imbalance. Only the Classla model with Llama augmentation with temperature = 2 was able to perform better. This shows the importance of having more complex and lengthy texts for training and the impact of the classification model chosen to fine-tune.

From these observations, the main takeaways from this study are:

- Better results can be achieved by balancing training sets using augmentation and keeping those sets balanced as more augmentation is added.
- Significant performance increases can be obtained using an LLM to generate augmentation data from the original data set.

- Increasing the temperature setting of the generating LLM can produce generated texts which are longer and more complex, improving performance over shorter texts that are closer in size and complexity to the original texts.
- Models trained on generated data have different levels of maximum performance. This study showed that by testing multiple models and varying augmentation levels, we could find the best combination of generator, temperature, and classifier to produce the best highest performance.

Using a higher temperature when generating texts from an LLM, specifically Llama, produced the best classification performance. The added benefit of using the higher temperature was that the generated responses were more diverse, allowing the model to continue improving, even at higher augmentation levels.

7 Future Work

This study was limited by the processing power needed to train and compare models. As such, we chose a limited number of temperatures to test on each model and limited the augmentation amount to 100x at its maximum. Even so, it took many weeks of round-the-clock fine-tuning to perform the testing. Several models using Llama for augmentation continued to rise in performance up to 100x. Continuing the augmentation to see the ultimate performance the model could achieve would be an important metric to record. Dynamic testing methods could also be used to limit unnecessary computation by pruning out models that are not promising. In future work, there is a need to test more combinations of factors in augmentation generation to exercise sufficient control and see if temperature was the only factor in achieving higher performance.

Further work should also be done to determine augmentation methods that improve performance on low-entropy data sets by identifying ways of enriching the minority classes of the training data for the model, resulting in more varied minority label samples and, therefore, higher performance.

Temperature variation in generated texts significantly altered fine-tuned model performance, and a combination of different temperatures or a combination of generating LLMs should be investigated further to see if there is a tighter correlation between complexity and downstream classifier performance.

Determining how much data is needed to maximize performance is traditionally complex to test empirically because a typical data set is obtained from a study where the number of instances produced is not a design feature. An alternative would be to use larger data sets trained on a model, sample the data, and augment the smaller set by adding data from the remaining data set to find where performance matches the full data set performance. Sampling the data requires stratification to ensure a representative label distribution for training.

Another issue to address is that the data may not have samples representative of the entire domain. The data set may contain responses that are only a subset of the entire domain space, leaving a gap in learning for part of the problem space. Understanding the degree of variance required in student responses could lead to a more generalized model that proves accurate across the entirety of the domain.

Another technique that could be employed in maximizing completion complexity and length and potentially expand the knowledge of a downstream classifier model to become more generalizable would be applying prompt engineering to the generator's input.

Finally, downstream tasks other than binary or multi-class classifiers should be tested using LLM-generated training data to determine if performance can increase for those particular tasks.

Declarations

- **Funding** The assessment project described in this article was funded, in part, by the Institute for Education Sciences, U.S. Department of Education (Grant R305G050091 and Grant R305F100007). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

- **Conflicts of interest/Competing interests** Not applicable
- **Availability of data and material** Data is currently protected under IRB rules.
- **Code availability** Not currently available for distribution.

References

Achieve, Inc (2013). *Next Generation Science Standards.* Washington, D.C: National Academies Press.

Azevedo, R., Johnson, A., Chauncey, A., Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with metacognitive tools. *New science of learning* (pp. 225–247). Springer.

Balkus, S.V., & Yan, D. (2023, August). Improving short text classification with augmented data using GPT-3. *Natural Language Engineering*, 1–30, https://doi.org/10.1017/s1351324923000438 Retrieved from http://dx.doi.org/10.1017/S1351324923000438

Basu, S., McElhaney, K.W., Rachmatullah, A., Hutchins, N., Biswas, G., Chiu, J. (2022). Promoting computational thinking through science-engineering integration using computational modeling. *Proceedings of the 16th International Conference of the Learning Sciences (ICLS).*

Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., 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*, 1–16,

Bayer, M., Kaufhold, M.-A., Reuter, C. (2021). A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*, ,

Biswas, G., Segedy, J.R., Bunchongchit, K. (2016). 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,

Britt, M.A., Rouet, J.-F., Durik, A.M. (2017). *Literacy beyond text comprehension: A theory of purposeful reading*. Routledge.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, ,

Chen, J., Tam, D., Raffel, C., Bansal, M., Yang, D. (2021). *An empirical survey of data augmentation for limited data learning in NLP*. Retrieved from https://arxiv.org/abs/2106.07499

Cho, Y., & Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing*, *37*, 25–38,

Cochran, K., Cohn, C., Hastings, P. (2023). Improving NLP model performance on small educational data sets using self-augmentation. *Proceedings of the 15th International Conference on Computer Supported Education.*

Cochran, K., Cohn, C., Hastings, P., Rouet, J.-F. (2023). Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. *Proceedings of the 24th International Conference on Artificial Intelligence in Education* (pp. 217–228).

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., Hastings, P. (2022). Improving automated evaluation of formative assessments with text data augmentation. *International Conference on Artificial Intelligence in Education* (pp. 390–401).

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, ,

Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E. (2021). A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, ,

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., Lester, J. (2021, 10). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education*, *31*, , https://doi.org/10.1007/s40593-020-00220-4

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A. (2020). Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, ,

Hastings, P., Hughes, S., Britt, A., Blaum, D., Wallace, P. (2014). Toward automatic inference of causal structure in student essays. *International Conference on Intelligent Tutoring Systems* (pp. 266–271).

HuggingFace (n.d.). *jeveuxaider/activity-classifier.* Retrieved from https://huggingface.co/jeveuxaider/activity-classifier

Hutchins, N.M., Basu, S., McElhaney, K.W., Chiu, J.L., Fick, S.J., Zhang, N., Biswas, G. (2021). Coherence across conceptual and computational representations of students' scientific models. *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021.*

Käser, T., & Schwartz, D.L. (2020). Modeling and analyzing inquiry strategies in open-ended learning environments. *International Journal of Artificial Intelligence in Education*, *30*(3), 504–535,

Kumar, A., Sharma, R., Bedi, P. (2024). Towards optimal nlp solutions: Analyzing gpt and llama-2 models across model scale, dataset size, and task diversity. *Engineering, Technology & Applied Science Research*, *14*(3), 14219–14224,

Kuzman, T., Mozetič, I., Ljubešić, N. (2023). Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149–1175,

Liu, P., Wang, X., Xiang, C., Meng, W. (2020). A survey of text data augmentation. *2020 International Conference on Computer Communication and Network Security (CCNS)* (pp. 191–195).

Luckin, R., & du Boulay, B. (2016). Reflections on the Ecolab and the Zone of Proximal Development. *International Journal of Artificial Intelligence in Education*, *26*(1), 416–430,

McElhaney, K.W., Zhang, N., Basu, S., McBride, E., Biswas, G., Chiu, J. (2020). Using computational modeling to integrate science and engineering curricular activities. M. Gresalfi & I. Horn (Eds.), *The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020* (Vol. 3).

Meta (n.d.). *Introducing Meta Llama 3: The most capable openly available LLM to date.* Retrieved from https://ai.meta.com/blog/meta-llama-3/

OECD (2021). *21st-century readers*. PISA, OECD Publishing. (https://www.oecd-ilibrary.org/content/publication/a83d84cb-en)

Piedboeuf, F., & Langlais, P. (2024). Data augmentation is dead, long live data augmentation. *arXiv preprint arXiv:2402.14895*, ,

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

Schwartz, R., & Stanovsky, G. (2022). On the limitations of dataset balancing: The lost battle against spurious correlations. *arXiv preprint arXiv:2204.12708*, ,

Seo, J.-W., Jung, H.-G., Lee, S.-W. (2021). Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning. *Neural Networks*, *138*, 140-149, https://doi.org/https://doi.org/10.1016/j.neunet.2021.02.007 Retrieved from https://www.sciencedirect.com/science/article/pii/S0893608021000496

Shorten, C., & Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48,

Shorten, C., Khoshgoftaar, T.M., Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*(1), 1–34,

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, ,

van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology*, *93*(3), 521,

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788,

Wikipedia (n.d.). *Gpt-3.* Retrieved from https://en.wikipedia.org/wiki/GPT-3

Winne, P.H., & Hadwin, A.F. (2013). nStudy: Tracing and supporting self-regulated learning in the internet. *International handbook of metacognition and learning technologies* (pp. 293–308). Springer.

Wu, L., Xie, P., Zhou, J., Zhang, M., Ma, C., Xu, G., Zhang, M. (2022). Self-augmentation for named entity recognition with meta reweighting. *arXiv preprint arXiv:2204.11406*, ,

Xia, C., Zhang, C., Zhang, J., Liang, T., Peng, H., Philip, S.Y. (2020). Low-shot learning in natural language processing. *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 185–189).

Yogatama, D., d'Autume, C.d.M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., ... others (2019). Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, ,

Zhang, N., Biswas, G., McElhaney, K.W., Basu, S., McBride, E., Chiu, J.L. (2020). Studying the interactions between science, engineering, and computational thinking in a learning-by-modeling environment. *International Conference on Artificial Intelligence in Education* (pp. 598–609).