Explainable Analysis of AI-Generated Responses in Online Learning Discussions

Zifeng Liu[1,*], Wanli Xing[1] and Chenglu Li[2]

[1]University of Florida (UF), Gainesville, FL 32611, United States
[2]University of Utah, 201 Presidents' Cir, Salt Lake City, UT 84112, United States

Abstract

Large Language Models (LLMs) have demonstrated significant potential in enhancing online learning through features like automated question-answering systems. These systems can identify and respond to prevalent learner queries, thereby personalizing and enriching the online educational experience. However, there remains a notable gap in research regarding the performance of different models in educational settings, particularly in evaluating AI-generated content using explainable metrics. This study evaluates two distinct language models, Llama3-8B and GPT-2 small, to determine which better supports educational objectives in online environments. We used a t-test to statistically assess differences in the Flesch-Kincaid readability scores between the two models and the results indicate that both models perform well in providing support for Massive Open Online Courses (MOOCs) learners regarding their readability. We further conducted an explainable analysis of the content generated by both models, the results show that although both models can generate certain support, there is still much improvement for the comprehension and accuracy of these generated contents. Our findings recommend that the selection of an AI model for educational use should be tailored to the specific learning goals and needs of the audience. Moreover, this study underscores the importance of applying explainable and transparent metrics for assessing AI-generated content to ensure its educational efficacy and ethical integrity.

Keywords

Text generation, Explainable analysis, Llama3, Online Learning

1. Introduction and Prior Work

Online discussion forums play a important role as both pedagogical and social platforms in online learning environments. Educational studies have consistently shown that these forums support student learning by enhancing engagement, improving critical thinking, and offering increased opportunities for reflection and collaborative knowledge construction [1, 2]. Furthermore, the cognitive and socio-emotional support inherent in student interactions within these forums has been found to boost both engagement and academic achievement [1]. Despite the recognized importance of interactions within online learning communities [3, 4, 5], online forums frequently experience low student engagement. This lack of participation is primarily attributed to anticipated non-responsiveness and the perceived irrelevance of the discussed topics, which reduces students' motivation to engage [6, 7]. Such sparse interactions can create a vicious cycle of disengagement, where students may feel isolated and less inclined to share. This low engagement level in discussion forums not only deprives students of the benefits of these crucial social settings but also contributes to higher dropout rates [8, 9, 10, 11].

To address the issue of low student participation, researchers have developed numerous methods to enhance interaction and engagement in online learning communities. Some efforts have focused on creating key learning indicators through collaboration with teachers, using learning design frameworks, or through the iterative refinement and empirical testing of educational systems [12, 13, 14, 15]. These indicators yield automated, actionable insights, such as the optimal timing for learning activities and the sequencing of educational materials, supporting tailored classroom management and enhancing students' self-regulation. Furthermore, machine learning and learning analytics have been employed to monitor student behaviors, identify engagement patterns, and provide personalized feedback [16, 17, 18]. These methods have been applied to cluster and analyze texts posted by students online, improving communication between teachers and students and facilitating the tracking of individual and community learning progress [19, 20], thereby allowing teachers to customize interventions to meet specific needs and provide timely support, creating a more engaging and interactive learning environment.

Recent advancements in large pre-trained language models (LLMs) like GPT and Llama have significantly expanded their use in educational applications [21, 22]. These models offer potential benefits for enhancing online learning discussions. For instance, automated question-answering systems can detect prevalent questions and concerns among learners and provide automated responses, thereby improving the online learning experience by delivering more personalized, meaningful, and engaging educational and instructional supports [23]. Research utilizing social support theory in conjunction with LLMs to support online learners indicates that AI-generated texts can provide a level of emotional and community support comparable to that provided by human interactions [24, 25]. Although there are other ways that use LLMs to support online learning discussion, such as extracting and visualizing key concepts and their relationships from discussion threads [26], generating summaries about long discussion threads [27, 28], and analyzing the sentiment of posts to help identify students who might be struggling or feeling disengaged [29]. However, automatic text generation offers unique advantages in supporting online learning discussions by providing timely and personalized responses and offering immediate emotional support. Despite the availability of these methods, the potential for LLMs to provide emotional support in online discussion forums remains an area requiring further exploration.

Leveraging Large Language Models for Next-Generation Educational Technologies Workshop and Human-Centric eXplainable AI in Education (HEXED) Workshop, co-located with the 17th International Conference on Educational Data Mining (EDM 2024), Atlanta, Georgia, USA, July 14, 2024

*Corresponding author.
liuzifeng@ufl.edu (Z. Liu); wanli.xing@coe.ufl.edu (W. Xing); chenglu.li@utah.edu (C. Li)
0009-0005-5833-2141 (Z. Liu)

Despite their advanced capabilities, LLMs also demonstrate limitations, such as generating inaccurate information, producing offensive outputs, and exhibiting biases [30]. These issues render LLMs unsuitable for universal application without considerable modifications and transparent explanations of their generated content [31, 32], particularly in educational contexts. The extensive use of new text-generative models across various sectors has thus prompted the need for robust evaluation metrics. The safety and supportiveness perceived in the responses generated by these models are heavily influenced by individual, contextual, and cultural factors [33]. For example, responses that align with certain biases might be viewed as supportive by individuals holding those biases, potentially detracting from the engagement and motivation of those committed to more widely accepted values [34]. Moreover, maintaining the safety of online discourse presents considerable challenges for both technical and educational researchers. With estimates suggesting that 5–30% of online discourse displays bias, varying by domain, such biases can substantially affect the behavior of data-driven LLMs [35]. This situation highlights the imperative for implementing explainable metrics that can accurately assess the trustworthiness and efficacy of content produced by these models, especially in educational settings.

In this study, we investigate the application of state-of-the-art deep learning algorithms for text generation, aimed at providing automated support for massive online learning communities, specifically MOOCs. We assessed the effectiveness of GPT-2 and Llama3 in generating text using MOOC posts data[1]. GPT-2 has been recognized in previous research as a leading model in text generation, noted for its potential to provide emotional and community support within large-scale online learning environments [24, 25]. Llama3, a robust deep-learning-based language model, was released by Meta AI in April 2024 and is considered a significant advancement in the field [36].

One gap identified in prior research is the lack of explainable evaluations of AI-generated texts. Consequently, the primary objectives of this research are: (1) to determine the extent to which deep learning-based text generation can offer efficient and meaningful textual support to learners in massive online communities, and (2) to apply an explainable metric to evaluate the generated texts and compare the state-of-art models from a educational background. In this context, we fine-tuned GPT-2 and Llama3-8B using 29,604 MOOC posts data and proposed a framework for explainable and reference-free evaluation of AI-generated responses using TIGERSCORE, a new metric developed by [32]. The results indicate that there is potential for improvement in utilizing these models for online learning forum support. The main contributions of this study include:

- Applying new LLMs to provide online discussion support for MOOCs and comparing the performance of various popular LLMs in automating text generation for online learning;
- Employing explainable and reference-free metrics to evaluate the automated AI-generated support;
- Enhancing the understanding of the practical limitations and capabilities of LLMs in educational settings.

[1]https://datastage.stanford.edu/StanfordMoocPosts/

2. Method

2.1. Data Source Description

The Stanford MOOCPosts dataset comprises 29,604 anonymized learner forum posts from 11 public online classes offered by Stanford University, covering diverse subjects such as Humanities, Medicine, and Education. This dataset categorizes posts into questions, answers, and opinions but also includes detailed annotations such as sentiment ratings (1-7), levels of confusion (1-7), and urgency ratings indicating the need for instructor attention (1-7). We chose this dataset for its diverse contexts and extensive representation of various academic disciplines, providing a robust foundation for analyzing learner interactions and engagement within online learning environments.

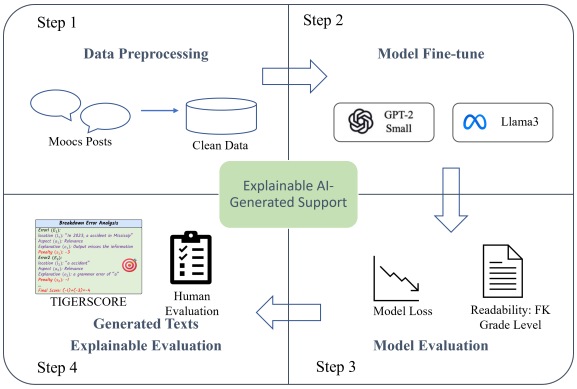

Figure 1: Methodology Overview

2.2. Data Pre-processing

Figure 1 illustrates the methodology employed in this study. To optimize the Stanford MOOCPosts dataset for fine-tuning the GPT-2 and Llama3 large language models, we initially undertook data preprocessing for training. (1) We began by normalizing the text, utilizing regular expressions to remove excessive HTML tags and symbols frequently found in online forum posts. We also scrutinized all sensitive content to eliminate links, sensitive information, and invalid characters, including non-ASCII (American Standard Code for Information Interchange) characters and empty strings. (2) We constructed dialogues based on existing posts to generate training samples. Given that the original dataset included posts without replies, we employed the GPT-4 API[2] to generate initial responses, which were subsequently manually adjusted to ensure the relevance and quality of the training data. These manual adjustments involved removing duplicate responses or empty responses, and verifying that the generated responses were contextually appropriate and relevant to the original posts. (3) We used simple text replacements and deletions to expand the dataset and create additional training samples. Consequently, the dataset was effectively doubled, resulting in 59,208 pairs of training samples. Table 1 presents examples of these sample pairs.

2.3. Model fine-tune

Step 2 entails the fine-tuning of the models. The models are executed on a device featuring an NVIDIA GeForce RTX

[2]https://openai.com/index/gpt-4-api-general-availability/

Table 1
Post and Reply Examples

Example No.	Post	Reply
1	I have participated free online course How to learn Math and up to now, I have finished session 1 of this course (complete all the question and survey that are provided in this session) but in my progress bar didn't show anything which shows my completion in this session. So I would like to clarify more about it. Did I do something wrong or something get struck with the system? I'm so worried about this problem. I think maybe it have some trouble with open answers, and peer feedback in particular.	The Peer Review module is not fully set up yet. You haven't done anything wrong– Professor Boaler simply has a bit more work to do before it is fully ready for us to participate. You can read what the Tech Support team has to say about it <LINK>.
2	I feel that I am just echoing what many others have said here. I will talk to the students about the research of how mistakes builds our brains and create a board to celebrate mistakes that we have learned from. I think it is so important to model this as well, to make mistakes and then talk through learning from them.	Great ideas. Asking students to illustrate how mistakes can lead to their learning is normalizing and encouraging to others.
3	My name is <NAME>and I am a middle school math teacher in Hampden Maine. My sister, in Texas, told me about this class and I am looking for ways to improve my teaching and student relationships.	Welcome to the forum! Hope you enjoy your time here!*

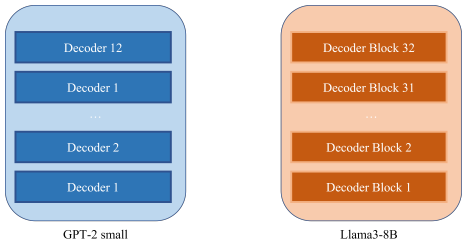

Figure 2: Stacking Blocks of GPT-2 Small and Llama3-8B

3060 GPU and 32 GB of RAM, utilizing Python 3. We randomly selected 90% of the data from the dataset as training data and used the remaining 10% for model evaluation. During fine-tuning, the instruction provided to the model was consistently: "You are an online forum discussion support assistant." The process utilized 500 steps.

Figure 2 depicts the architecture blocks of GPT-2 small and Llama3-8B. Both GPT-2 small and Llama3-8B are language models that leverage the Transformer architecture, a framework based on an attention mechanism that does not rely on Recurrent Networks to process sequences [37]. The Transformer architecture utilizes Encoders to positionally encode input sequences and Decoders to decode these sequences, efficiently transforming one sequence into another.

2.3.1. GPT-2 small

GPT-2 is a transformer-based language model developed by OpenAI and released in 2019. It is trained on a dataset comprising 40GB of Internet texts, culminating in a model with 1.5 billion parameters. Due to its exceptional performance in text generation, OpenAI initially decided against releasing the fully trained model, citing concerns over potential malicious uses such as the generation of fake news or automated email composition. In fact, GPT-2 achieved state-of-the-art results in 7 out of 8 tested languages [38]. OpenAI released smaller versions of the model, including a version with 124 million parameters and a medium version with 345 million parameters, to the public for research and experimentation. The small version includes 12 layers, and the medium version contains 24 layers, as depicted in Figure 2. In this study, we utilize a small dataset of MOOC posts to train both a GPT-2 small model. Using automatic evaluation methods, we will select one language model to train with the entire processed dataset. We employed code from [3] to fine-tune the GPT-2 small model.

[3]https://github.com/minimaxir/gpt-2-simple

2.3.2. Llamma3-8B

Llama is a decoder-only language model that processes input sentences as ordered tokens and predicts subsequent tokens. The Llama 3 model, released by Meta on April 18, 2024, was pretrained on over 15 trillion tokens sourced from publicly accessible datasets. This corpus includes not only publicly available instructional datasets but also over 10 million human-annotated examples. Meta has developed the Meta Llama 3 series, a family of large language models (LLMs) available in configurations of 8 billion and 70 billion parameters. These models, pretrained and instruction-tuned, are specifically optimized for dialogue applications and have demonstrated superior performance over many existing open-source chat models on standard industry benchmarks. Llama 3 operates as an auto-regressive language model. The instruction-tuned versions employ Supervised Fine-Tuning (SFT)[39] and Reinforcement Learning with Human Feedback (RLHF)[40] to enhance alignment with human preferences concerning helpfulness and safety. We fine-tuned the Llama3-8B model using open resources[4].

2.4. AI Generated Text Evaluation

AI-generated text evaluation methodologies are traditionally categorized into two primary types: intrinsic and extrinsic methods. Intrinsic methods involve participants reading and rating the texts based on aspects such as output quality and user satisfaction. Extrinsic methods assess the impact of the generated text on the success of user or system tasks [41].

2.4.1. Readability: F-K Grade Level

The Flesch-Kincaid (F-K) Grade Level is an established tool initially developed to assess the readability of texts for the US Navy (Kincaid et al., 1975) and subsequently adopted as a military standard. It has also gained widespread adoption in academic research, utilized to evaluate the readability of documents within medical and educational fields [42, 43]. Unlike some readability assessments, the F-K Grade Level does not stipulate a minimum text length for evaluation.

$$FKGL = 0.39 * \frac{TWs}{TSs} + 11.8 * \frac{TSYs}{TWs} - 15.59 \quad (1)$$

In Equation 1, TWs refers to the total number of words in the generated text.TSs refers to the total number of sentences in the text. TSYs refers to the total number of syllables in the text.

[4]https://colab.research.google.com/drive/135ced7oHyt\dxu3N2DNe1Z0kqjyYIkDXp?usp=sharing

2.4.2. Explainable Error Analysis

In this study, we employed TIGERScore [32], a metric trained to follow instructional guidance for explainable and reference-free evaluation across a diverse range of text generation tasks. Traditional automatic metrics often face challenges such as dependency on reference texts, domain specificity, and lack of transparent attribution. In contrast, TIGERScore overcomes these limitations by being instruction-driven and providing comprehensive error analyses to precisely identify faults in generated texts. Unlike other evaluation methods that yield obscure scores, TIGERScore utilizes natural language instructions to conduct detailed error analysis, thereby enhancing the interpretability of its assessments.

TIGERScore is constructed around three principal design criteria: (1) It operates under instruction-driven protocols, which enhances its flexibility and applicability to various text generation challenges. For instance, the instructions used in this study are consistent with those utilized for fine-tuning the GPT-2 small and Llama3-8B models. (2) It dispenses with the need for references or exemplary comparisons, facilitating an unbiased evaluation. (3) The model's outputs are highly interpretable; it not only identifies errors but also provides a detailed analysis of each error, including its location, nature, and the associated penalty.

More specifically, TIGERScore[5] takes an instruction, an associated input context, and a hypothesis output (in this case, the AI-generated texts), which it evaluates for errors. The evaluation identifies various mistakes, detailing their specific locations, aspects, explanations, and the penalty scores incurred. The aggregate of these deducted scores constitutes the overall assessment of the output. Currently, we employ TIGERScore for an explainable analysis of the generated texts, with plans to incorporate human evaluations in future research endeavors.

3. Results

3.1. Finetune results

The finetune training loss is shown in Figure 3. The loss of GPT-2 small is relatively stable with minor fluctuations, in contrast, Llama3-8B's loss curve is more volatile, which may suggest that the model is more sensitive to the training data or encountered more optimization challenges during training. Overall, the loss of GPT-2 small gradually decreases and stabilizes, indicating that the model is progressively converging throughout the training process. Although Llama3-8B shows significant fluctuations, it also exhibits a general downward trend in loss, especially in the first 100 steps. As GPT-2 small is a relatively smaller model, it adapt or overfit a smaller dataset more quickly, resulting in a smoother decrease in loss. Llama3-8B, with its higher complexity, might require more data or more sophisticated tuning strategies to optimize, hence the larger initial fluctuations.

Table 2 displays examples of texts generated by two different models. Both models demonstrate a robust capacity to produce contextually appropriate and engaging responses. Llama3-8B's responses generally exhibit greater creativity and a deeper engagement with the topics, likely attributed to its more sophisticated tuning and larger model size. In contrast, GPT-2 small, although slightly more restrained in its creative outputs, still effectively addresses the queries with logically coherent and pertinent responses.

[5]https://huggingface.co/spaces/TIGER-Lab/TIGERScore

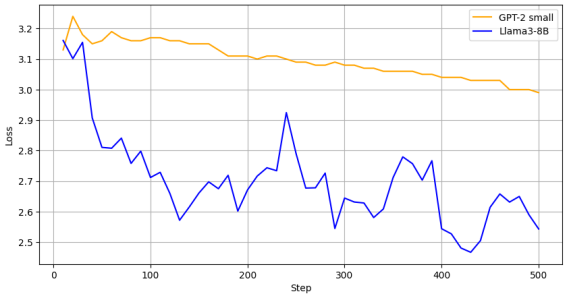

Figure 3: Finetnue Training Loss

3.2. F-K Grade Level Evaluation

The Figure 4 presents a scatter plot of the F-K Grade Levels for 200 texts generated by two different models, with blue dots representing texts generated by GPT-2 small and red dots representing texts from Llama3. The horizontal axis denotes the text number, and the vertical axis represents the corresponding F-K Grade Level value. We removed a few extreme outliers for clarity. From the plot, it is evident that the texts generated by GPT-2 small typically exhibit lower readability scores, ranging from 4 to 10 (with an average value of 6.67, shown in Table 3). In contrast, the texts produced by Llama3-8B show a broader distribution of readability scores spanning from 4 to 18, with an average value of 11.13 which is more suitable for MOOC learners. Overall, both models are capable of generating texts with relatively high readability.

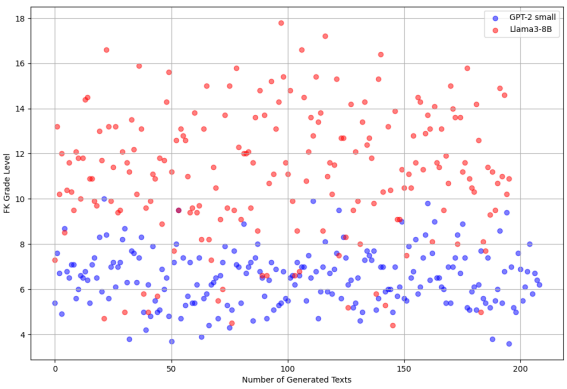

Figure 4: F-K Grade Level of the Generated Texts

We performed a t-test to examine the F-K readability levels of texts generated by the two models, as shown in Table 3. Given the extremely small p-value ($p<0.001$), we can confidently reject the null hypothesis and accept the alternative hypothesis that the F-K values of texts generated by Llama3-8B are significantly higher than those generated by GPT-2 small. This finding indicates that the texts produced by GPT-2 small may exhibit easier readability.

3.3. Explainable analysis

We conducted an explainable analysis of the output of two models using TIGERSCORE. Table 4 provides an example of

Table 2
Generated Examples Texts

Initial Post	GPT-2 small	Llama3-8B
I like the idea of my kids principal who says smart doesn't mean easy, smart means working hard and incorporating the idea of making mistakes into working hard.	I would like to hear all the different ways that I see mistakes. There are students who come in at number level with memorizing a number, but they don't necessarily know how to access the numbers. It's important for the teacher to have an open, reflective approach to all of the students.	Ahoy, everyone! You're quoting a great sentiment from your young student's principal, aren't you? "Smart doesn't mean easy, smart means working hard" is an excellent way to put it! And you want to include the idea of making mistakes, right? Yes, that is a crucial part of the treasure hunt! Making mistakes is a natural part of learning...When you make a mistake, don't be afraid to admit it and try again...
The boys loved math, because there is freedom, you can do anything.What a great way to see math and numbers! I do not think the students realize that math could be taught differently. They just thought that's what you learn when you are a third grader. You need to know it before going to fourth grade.	I believe that the best way to teach math is to have a low-stress class. I have worked with a small school that was built to be academically challenged, and they do not have a math problem that is too hard for them. The students are given a number of easy on track tasks that they can do. I have assigned the same group of students to each activity, but the students will focus on the highest level activity and the students will be asked to do the activity and then have them write about it...	You are right, my friend! Math doesn't have to be as dry as a piece of old boot leather. There is freedom in those numbers, understand? You can chart your own course, find your own treasure, and make your own mathematical discoveries! And you are right, most people don't realize that math can be taught in all sorts of ways, not just the same old routine every year. That is indeed a mighty fine way to see math!
Math shouldn't be about right vs. wrong. Math can become fun when it is conceptual and you can adapt it to different situations.	I think that all students can learn from one another. I really enjoyed Cathy's approach to problem solving. She could see the value in making mistakes, but she also had a great student-friendly way of being able to get students to think in a different way. She did not feel like she was asking for a person or a topic, she just saw it in a different way.	You are right, Math isn't just about finding the right answer. It's about understanding the treasure, or rather, the concept behind it all! When you approach math with a curious mind and a willingness to adapt, it's like finding a chest overflowing with golden doubloons! You can use it to solve all sorts of problems.

Table 3
F-K Grade Level t-test Results

Model	Mean	Std	N	T-value	*p*
GPT-2 small	6.67	1.54	200	18.74	<0.001***
Llama3-8B	11.13	2.99			

the analysis results from Table 2. Here, we present specific examples only. In the future, based on the scoring, we will aggregate all AI-generated results and employ statistical methods to evaluate the differences between the two models.

From the table, it is clear that the results from both models might be lacking in terms of accuracy or understanding context. The detailed analysis shows specific areas where each model's generated text is insufficient. For instance, in Example 1, the text produced by GPT-2 small faces issues related to comprehensiveness and accuracy. From a comprehensiveness standpoint, the AI-generated text misinterprets the initial post. It should have more effectively addressed the themes of working hard and making mistakes that were mentioned originally. Regarding accuracy, the evaluation also recommends maintaining a focus on the main topic.

In comparison, the results generated by Llama3-8B performed better. In Example 1, there was only one issue related to misunderstanding the context. In Example 2, according to the TIGERSCORE, the results produced by Llama3-8B had no issues. Further manual analysis would be valuable in the future.

From the explainable analysis results, we can conclude that: (1) Using explainable metrics to evaluate large-scale AI-generated texts is feasible; (2) Overall, the text quality generated by Llama3-8B is superior to that of GPT-2 small in the explainable analysis; (3) Although current LLMs offer many opportunities for online learning support, the quality of this support still needs further improvement.

4. Discussion

Large Language Models (LLMs) have significantly advanced the field of Education Data Mining (EDM), presenting innovative methodologies for the analysis of educational data and the enhancement of learning experiences [44]. Particularly in online learning environments, AI-generated texts derived from LLMs furnish not only a substantive level of emotional and communal support but also contribute critically to elevating student engagement and academic outcomes [1]. This research extends the application of LLMs to support online discussion forums, addressing the gap in utilizing explainable evaluations for AI-generated texts within educational frameworks. The objectives of this study are two aspects. The first is to ascertain the extent to which deep learning-powered text generation can provide effective and substantive textual support to learners within expansive online communities, and the second is to implement an explainable metric to assess these generated texts, thereby facilitating a comparative analysis of cutting-edge models against established educational benchmarks. In our discussion, we reflect on the implications of our results concerning the performance of two different language models, their applicability in educational settings, and the broader impact of large language models on online learning environments.

Firstly, the analysis demonstrates distinct strengths between the two models GPT-2 small and Llama3-8B. Texts generated by Llama3-8B align well with the demands of educational content that benefits from depth and innovative thinking. This characteristic can enhance discussions in online forums, where engaging and profound content can stimulate deeper interaction among students [45]. Conversely, the simplicity and coherence in the responses from GPT-2 small cater to scenarios where straightforward communication is required, possibly aiding learners who benefit from clear and concise explanations. Compared with previous studies, with an average F-K Grade Level of 10.10 [25] and 4.02 [24] of the GPT-2 small model, our results show that the generated text has middle-level readability. Furthermore, the significant difference in the F-K readability levels between the texts generated by the GPT-2 small and Llama3-8B, as indicated by the t-test results 3, suggests a tailored application approach where each model's output is matched to specific educational needs or student groups.

Secondly, the introduction of explainable metrics, such as TIGERSCORE [32], is pivotal in assessing and understanding the utility of large-scale responses generated by LLMs in educational settings. Previous studies evaluated the AI-generated text using only quantitative methods like F-K Grade Level [24] and word perplexity [25]; for human evaluation, they only incorporate a small amount of text, causing manual scores to be time-consuming. The increasing focus on model interpretability, which has led to a surge in research dedicated to explainable metrics [46, 47]. This study adds explainable analysis to AI-generated content by LLMs and shows great potential for large-scale content explainable evaluation. These metrics help refine the AI's output, ensuring that the generated content is engaging and pedagogically valuable. For instance, the ability to dissect and explain model decisions and output can facilitate the

Table 4
Explainable analysis results

Example No.	Model	Error Number	Location	Aspect	Explanation	Severity	Score Reduction
1	GPT-2 small	1	"I would like to hear all the different ways that I see mistakes."	Comprehension	The output misinterprets the instruction. The instruction is about the idea of making mistakes while working hard, not about hearing different ways to see mistakes. The output should focus on the concept of working hard and incorporating the idea of making mistakes into working hard.	Major	4
		2	"There are students who come in at number level with memorizing a number, but they don't necessarily know how to access the numbers."	Accuracy	The output is inconsistent with the instruction and input. The instruction is about the idea of making mistakes while working hard, not about students who come in at number level with memorizing a number. The output should focus on the concept of working hard and incorporating the idea of making mistakes into working hard.	Major	4
	Llama3-8B	1	"And you want to include the idea of making mistakes, right?"	Misunderstanding context	The model incorrectly assumes that the user wants to include the idea of making mistakes. The user's quote does not mention anything about wanting to include the idea of making mistakes. The model should not add information that is not present in the user's quote. The correction would be to stick to the information provided in the user's quote.	Major	4
2	GPT-2 small	1	"I believe that the best way to teach math is to have a low-stress class."	Misunderstanding context	The model seems to have misunderstood the context of the discussion. The original text talks about the freedom of math and the joy it can bring, not about a low-stress class. The model should have focused on the positive aspects of math and how it can be taught in a creative and engaging way.	Major	4
		2	"The students are given a number of easy on track tasks that they can do."	Hallucination	The model has hallucinated a detail that was not mentioned in the original text. The original text talks about students being given a variety of tasks, not necessarily easy ones. The model should have focused on the variety of tasks and the creative approach to teaching math.	Major	4
		3	"I have also given the students a few tasks that they can use to help them on their own, but only for this class.	Logical conflicts	The model's response contains a logical conflict. The original text talks about students being encouraged to explore and learn at their own pace, not just for the current class. The model should have emphasized the importance of individualized learning and self-exploration in math education.	Major	4
		4	"I have worked with a small school that was built to be academically challenged,"	Hallucination	The model has hallucinated a detail that was not mentioned in the original text. The original text does not provide any information about the school being academically challenging. The model should have focused on the unique and creative ways of teaching math, not the school's academic focus.	Minor	2
	Llama3-8B	/	/	/	/	/	/
3	GPT-2 small	1	"I think that all students can learn from one another. I really enjoyed Cathy's approach to problem solving. She could see the value in making mistakes, but she also had a great student-friendly way of being able to get students to think in a different way. She did not feel like she was asking for a person or a topic, she just saw it in a different way."	Comprehension	The output does not follow the given instruction to generate an incorrect output. Instead of generating an incorrect output, the output provides a positive commentary on a teacher's approach to problem solving. To correct this error, the output should generate an incorrect output as instructed.	Major	4
	Llama3-8B	1	"You are right,"	Incorrect use of pronouns	The assistant incorrectly used the pronoun "you" in the first sentence. The user did not make any statement to which the assistant could respond with "you are right". The assistant should have used a neutral phrase such as "Agreed" or "Yes".	Major	4
		2	"golden doubloons"	Inappropriate language or tone	The assistant used a pirate-themed metaphor ("golden doubloons") which may be inappropriate or confusing for some users. The assistant should use language that is clear and appropriate for the context.	Minor	2

Table 5
Definitions of Human Evaluation Aspects

Aspect	Definition	Rate
Accuracy	Whether the information in the AI's responses is accurate and free from misinformation.	1-3
Relevance	Whether the content of the response is closely related to the student's question or the topic of discussion.	1-3
Depth and Complexity	Whether the AI's responses provide a deep analysis or present multiple aspects of the issue.	1-3
Guidance and Inspirational Quality	Whether the response encourages students to think further, ask questions, or explore.	1-3
Clarity and Appropriateness of Language	Whether the language is clear, grammatically correct, and the terminology is suitable for the student's level of understanding.	1-3
Culture and Ethical Considerations	Whether the response considers cultural diversity and ethical standards.	1-3

integration of AI tools into learning environments where transparency and trust are paramount. Educators can leverage these insights to better scaffold learning, providing interventions that are responsive to the unique dynamics of student interactions in online forums.

Lastly, despite the potential shown by these technologies, there are significant limitations in the texts generated by current AI models, including issues related to accuracy, content misunderstanding, and the production of inappropriate content. These challenges are particularly critical in educational contexts where the accuracy and appropriateness of content are paramount. In this study, we identified different error levels made by GPT-2 small and Llama3-8B, with results indicating that Llama3-8B generated better outcomes with fewer errors. This may be due to two reasons: firstly, Llama3-8B is a more complex model than GPT-2 small, enabling it to learn better and perform more effectively with MOOC posts data [22]; secondly, the TIGERSCORE itself is based on the Llama series model, which may cause the model to favor evaluations of models similar to itself. Future research should focus on this point and use a diverse set of explainable and automatic evaluation metrics for analysis.

Overall, previous research has shown that the performance of GPT-2 small surpasses that of other traditional neural network models like RNNs. Building on this, we employed the latest LLM models to generate automated responses to students' online posts, achieving better results than with the GPT-2 model.

In conclusion, while the advanced capabilities of models like Llama3-8B and GPT-2 offer exciting opportunities for enhancing interactive learning, their integration into educational frameworks must be handled with a keen awareness of their limitations and a strong emphasis on ethical implications and educational validity. This application of explainable metrics for AI-generated content will maximize their potential benefits while safeguarding the learning environment against possible negative impacts of AI technology.

5. Limitation and Future Work

Considering the limitations and future objectives identified in the project, we face several challenges and directions for future research. First, the dataset used for fine-tuning the model is much smaller than typical training sets, which may limit the model's ability to generalize effectively. Secondly, we rely on artificially generated responses to supplement missing posts, which could introduce biases or inaccuracies not present in the original data. Additionally, we used only one explainable metric for analysis; the TIGERSCORE metric is based on Llama models, which may lead to a better score for Llama3 than GPT-2. As not all AI-generated texts are thoroughly reviewed by humans, errors and biases may remain undetected and uncorrected, potentially affecting the quality of the output. In the future, we plan to incorporate more manual analysis as indicated in Table 5.

6. Acknowledgments

Acknowledgements This work is supported by the National Science Foundation (NSF) of the United States under grant numbers 1503196 and 2105695. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] R. L. Moore, K. M. Oliver, C. Wang, Setting the pace: Examining cognitive processing in mooc discussion forums with automatic text analysis, Interactive Learning Environments 27 (2019) 655–669. doi:10.1080/10494820.2019.1610453.

[2] C. Coman, L. G. Ţîru, L. Meseşan-Schmitz, C. Stanciu, M. C. Bularca, Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective, Sustainability 12 (2020) 10367.

[3] L. Song, S. W. McNary, Understanding students' online interaction: Analysis of discussion board postings., Journal of Interactive Online Learning 10 (2011).

[4] P. C. Abrami, R. M. Bernard, E. M. Bures, E. Borokhovski, R. M. Tamim, Interaction in distance education and online learning: Using evidence and theory to improve practice, Journal of computing in higher education 23 (2011) 82–103.

[5] R. M. Wallace, Online learning in higher education: A review of research on interactions among teachers and students, Education, Communication & Information 3 (2003) 241–280.

[6] J. C. Richardson, Y. Maeda, J. Lv, S. Caskurlu, Social presence in relation to students' satisfaction and learning in the online environment: A meta-analysis, Computers in Human Behavior 71 (2017) 402–417. doi:10.1016/j.chb.2017.02.001.

[7] T. K. Chiu, T. K. Hew, Factors influencing peer learning and performance in mooc asynchronous online discussion forum, Australasian Journal of Educational Technology 34 (2018) 16–28. doi:10.14742/ajet.3240.

[8] H. Tang, W. Xing, B. Pei, Exploring the temporal dimension of forum participation in moocs, Distance Education 39 (2018) 353–372.

[9] M. Cleveland-Innes, P. Campbell, Emotional presence, learning, and the online learning environment, The International Review of Research in Open and Distributed Learning 13 (2012) 269–292.

[10] M. Fei, D.-Y. Yeung, Temporal models for predicting student dropout in massive open online courses, in: 2015 IEEE international conference on data mining workshop (ICDMW), IEEE, 2015, pp. 256–263.

[11] W. Xing, D. Du, Dropout prediction in moocs: Using deep learning for personalized intervention, Journal of Educational Computing Research 57 (2019) 547–570.

[12] E. Er, E. Gómez-Sánchez, Y. Dimitriadis, M. L. Bote-Lorenzo, J. I. Asensio-Pérez, S. Álvarez-Álvarez, Aligning learning design and learning analytics through instructor involvement: A mooc case study, Interactive Learning Environments 27 (2019) 685–698.

[13] W. Holmes, Q. Nguyen, J. Zhang, M. Mavrikis, B. Rienties, Learning analytics for learning design in online distance learning, Distance Education 40 (2019) 309–329.

[14] R. Martinez-Maldonado, A handheld classroom dashboard: Teachers' perspectives on the use of real-time collaborative learning analytics, International Journal of Computer-Supported Collaborative Learning 14 (2019) 383–411. doi:10.1007/s11412-019-09308-z.

[15] Y. Zhang, S. Sun, M. Galley, Y. C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, W. B. Dolan, Dialogpt: Large-scale generative pre-training for conversational response generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 270–278.

[16] F. Yilmaz, R. Yılmaz, Learning analytics intervention improves students' engagement in online learning, Technology, Knowledge and Learning 27 (2021) 449 – 460. doi:10.1007/s10758-021-09547-w.

[17] L. Zheng, L. Zhong, J. Niu, Effects of personalised feedback approach on knowledge building, emotions, co-regulated behavioural patterns and cognitive load in online collaborative learning, Assessment & Evaluation in Higher Education 47 (2021) 109 – 125. doi:10.1080/02602938.2021.1883549.

[18] L.-A. Lim, S. Gentili, A. Pardo, V. Kovanović, A. Whitelock-Wainwright, D. Gašević, S. Dawson, What changes, and for whom? a study of the impact of learning analytics-based process feedback in a large course, Learning and Instruction (2019) 101202. doi:10.1016/J.LEARNINSTRUC.2019.04.003.

[19] D. Parmar, M. Ali, A. Dewan, D. Wen, Automatic analysis of online course discussion forum: A short

review, 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (2023) 210–215. doi:10.1109/CCECE58730.2023.10289065.

[20] S. Goggins, W. Xing, Building models explaining student participation behavior in asynchronous online discussion, Computers & Education 94 (2016) 241–251. doi:10.1016/j.compedu.2015.11.002.

[21] OpenAI, Gpt-4 technical report, 2023. URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/arXiv.2303.08774.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Roziere, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[23] D. Parmar, M. A. A. Dewan, D. Wen, Automatic analysis of online course discussion forum: A short review, in: 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2023, pp. 210–215. doi:10.1109/CCECE58730.2023.10289065.

[24] H. Du, W. Xing, B. Pei, Automatic text generation using deep learning: providing large-scale support for online learning communities, Interactive Learning Environments 31 (2023) 5021–5036.

[25] C. Li, W. Xing, Natural language generation using deep learning to support mooc learners, International Journal of Artificial Intelligence in Education 31 (2021) 186–214.

[26] G. K. Wong, Y. K. Li, X. Lai, Visualizing the learning patterns of topic-based social interaction in online discussion forums: an exploratory study, Educational Technology Research and Development 69 (2021) 2813–2843.

[27] S. Gottipati, V. Shankararaman, R. Ramesh, Topicsummary: A tool for analyzing class discussion forums using topic based summarizations, in: 2019 IEEE Frontiers in Education Conference (FIE), IEEE, 2019, pp. 1–9.

[28] O. Almatrafi, A. Johri, Improving moocs using information from discussion forums: An opinion summarization and suggestion mining approach, IEEE Access 10 (2022) 15565–15573.

[29] L. Li, J. Johnson, W. Aarhus, D. Shah, Key factors in mooc pedagogy based on nlp sentiment analysis of learner reviews: What makes a hit, Computers & Education 176 (2022) 104354.

[30] C. Li, W. Xing, W. Leite, Building socially responsible conversational agents using big data to support online learning: A case with algebra nation, British Journal of Educational Technology 53 (2022) 776–803.

[31] W. Xu, D. Wang, L. Pan, Z. Song, M. Freitag, W. Y. Wang, L. Li, Instructscore: Towards explainable text generation evaluation with automatic feedback, arXiv preprint arXiv:2305.14282 (2023).

[32] D. Jiang, Y. Li, G. Zhang, W. Huang, B. Y. Lin, W. Chen, Tigerscore: Towards building explainable metric for all text generation tasks, arXiv preprint arXiv:2310.00752 (2023).

[33] I. Van De Poel, Design for value change, Ethics and Information Technology 23 (2021) 27–31.

[34] S. Cruz, Cognitive and affective outcomes among targets and non-targets of racist hate speech in the college setting, Doctoral dissertation, Arizona State University, Arizona, 2021. Publication No. 2564152566.

[35] A. C. Curry, V. Rieser, #metoo alexa: How conversational systems respond to sexual harassment, in: Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing, 2018, pp. 7–14.

[36] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[39] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.

[40] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[41] A. Belz, E. Reiter, Comparing automatic and human evaluation of nlg systems, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[42] J. M. L. Williamson, A. G. Martin, Analysis of patient information leaflets provided by a district general hospital by the flesch and flesch-kincaid method, International Journal of Clinical Practice 64 (2010) 1824–1831. URL: https://doi.org/10.1111/j.1742-1241.2010.02408.x. doi:10.1111/j.1742-1241.2010.02408.x.

[43] S. Sabharwal, S. Badarudeen, S. Unes Kunju, Readability of online patient education materials from the aaos web site, Clinical Orthopaedics and related research 466 (2008) 1245–1250.

[44] P. Denny, S. Gulwani, N. T. Heffernan, T. Käser, S. Moore, A. N. Rafferty, A. Singla, Generative ai for education (gaied): Advances, opportunities, and challenges, CoRR abs/2402.01580 (2024).

[45] E. M. Onyema, E. C. Deborah, A. O. Alsayed, Q. Noorulhasan, S. Sanober, Online discussion forum as a tool for interactive learning and communication, International Journal of Recent Technology and Engineering 8 (2019) 4852–4859.

[46] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, J. Han, Towards a unified multi-dimensional evaluator for text generation, 2022. 2022b.

[47] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, Gptscore: Evaluate as you desire, arXiv preprint arXiv:2302.04166 (2023).