JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(3), 8–33. https://doi.org/10.18608/jla.2025.8885

Exploring Fairness and Explainability in LLM-Generated Support for Online Learning Discussion Forums

Zifeng Liu[1], Wanli Xing[2], Xinyue Jiao[3] and Chenglu Li[4]

Abstract

Large language models (LLMs) hold significant potential to enhance online learning by automating responses to learner queries and offering personalized, scalable support. However, concerns about bias in LLM-generated responses present challenges to their ethical and equitable use in educational settings. This study explores fairness and explainability in LLM-generated replies within online discussion forums. Specifically, we fine-tuned three state-of-the-art LLMs (GPT-2, Gemma, and LLaMA) using both the original MOOC Posts dataset and a counterfactual version. We then analyzed the sentiment patterns of LLM-generated replies and compared them with human-generated responses. To quantify potential sentiment bias, we introduce absolute distributional sentiment divergence (ADSD) to measure disparities across sensitive attributes, with gender used as a case study. To mitigate bias and enhance transparency, we employed counterfactual fine-tuning by incorporating both factual and counterfactual data, and we used TIGERSCORE, a reference-free explainability metric, to assess response quality. Our findings reveal that LLM-generated responses are generally more neutral than human replies but exhibit varying degrees of sentiment bias across gender. Notably, counterfactual fine-tuning shows promise in reducing this bias, resulting in more balanced sentiment distributions. Additionally, explainability analysis indicates that while newer models (Gemma and LLaMA) outperform GPT-2 in response quality, gaps in accuracy and comprehension remain. This study advances the understanding of bias mitigation and fairness evaluation in LLM-generated educational support, contributing to the development of more equitable, transparent, and responsible AI-driven tools for online learning environments.

Notes for Practice

- Large language models (LLMs) generate automated responses in online learning forums that differ in sentiment from human replies, often appearing more neutral.
- We propose absolute distributional sentiment divergence (ADSD), a novel metric to evaluate and reduce sentiment bias across sensitive attributes such as gender.
- Counterfactual fine-tuning combining original and counterfactual data effectively reduces sentiment bias in LLM-generated responses.
- Using TIGERSCORE, a reference-free explainability metric, we find that Gemma and LLaMA produce higher-quality responses than GPT-2, though accuracy and comprehension challenges persist.

Keywords: Sentiment bias, text generation, fairness, explainable analysis, large language model, online learning.

Submitted: 29/01/2025 — **Accepted:** 12/07/2025 — **Published:** 03/09/2025

[1] *Email: liuzifeng@ufl.edu Address: College of Education, University of Florida, Gainesville, Florida, USA. ORCID iD: https://orcid.org/0009-0005-5833-2141*

[2] *Corresponding author Email: wanli.xing@coe.ufl.edu Address: College of Education, University of Florida, Gainesville, Florida, USA. ORCID iD: https://orcid.org/0000-0002-1446-889X*

[3] *Email: xj2320@nyu.edu Address: School of Culture, Education, and Human Development, New York University, New York, USA. ORCID iD: https://orcid.org/0000-0001-9496-609X*

[4] *Email: chenglu.li@utah.edu Address: Department of Educational Psychology, University of Utah, Salt Lake City, Utah, USA. ORCID iD: https://orcid.org/0000-0002-1782-0457*

1. Introduction

Online learning has become a vital component of contemporary higher education, with online discussion forums serving as crucial pedagogical and social platforms within these learning environments. Asynchronous discussion forums are one of the most common forms and are typically used to promote deep learning, collaborative learning, and the recording and assessment of student progress (Hew & Cheung, 2012). Despite the recognized importance of interactions within online learning communities (Abrami et al., 2011), online forums frequently suffer from low student engagement (Hew & Cheung, 2012; Dumford & Miller, 2018). This lack of participation is primarily attributed to students' expectations of non-responsiveness, limited feedback, and the perceived irrelevance of discussion topics. These factors collectively reduce students' motivation to engage. Consequently, low engagement not only prevents students from benefiting from these valuable social settings but also contributes to higher dropout rates (Tang et al., 2018; Xing & Du, 2019).

Recent advancements in large language models (LLMs) have presented potential solutions for addressing these challenges in educational settings (Baidoo-Anu & Ansah, 2023; Song et al., 2024; Hwang & Chen, 2023). Researchers have explored various ways to use LLMs to support online learning discussions, such as extracting and visualizing key concepts and their relationships from discussion threads (Wong et al., 2021) and generating summaries of lengthy conversations (Gottipati et al., 2019; Almatrafi & Johri, 2022). Among these approaches, the automatic text generation capabilities of LLMs offer unique advantages by providing timely, personalized responses and immediate feedback (H. Li et al., 2024; Naseer et al., 2024; Q. Wang et al., 2022). Large online learning forums, especially in MOOCs, often contain thousands of posts, making it difficult for instructors to respond to every query. LLMs can manage a large volume of interactions, ensuring timely responses for more students. Their ability to operate 24/7 also provides continuous support, especially during hours when instructors are unavailable (Yan et al., 2024; Parmar et al., 2023). For example, automated question-answering systems can identify common content-related queries and generate appropriate responses. This enhances the online learning experience by offering more personalized, meaningful, and engaging educational support (Zylich et al., 2020; Sahay et al., 2019).

From a theoretical perspective, social support theory emphasizes the importance of providing emotional support, informational support, and companionship to help individuals cope with stress and challenges (Shumaker & Brownell, 1984; Langford et al., 1997; Thoits, 1986). Researchers have found that socio-emotional interactions play a vital role in fostering student engagement in online discussions (Rovai, 2007). However, despite the growing use of LLMs in educational settings, limited research has explored their potential to support the emotional dimension of such interactions. While many studies have used sentiment analysis to identify students who may be struggling or feeling disengaged in online environments (L. Li et al., 2022; Hew et al., 2020; Du & Xing, 2023; Onan, 2021), most focus on detecting and classifying student posts, reviews, or comments rather than exploring how LLMs can actively provide socio-emotional support during interactions. To the best of our knowledge, only two studies have manually evaluated LLM-generated responses in online discussions and concluded that AI-generated texts can offer a certain degree of emotional support (Du et al., 2023; C. Li & Xing, 2021). However, there remains a lack of quantitative analysis of the sentiment patterns exhibited in these responses. To effectively utilize LLMs in supporting socio-emotional interactions, it is crucial to understand the sentiment patterns in their generated responses. Moreover, with the rapid development of powerful LLMs, new opportunities are emerging to leverage these models in online learning contexts. Yet, few studies have fully explored or taken advantage of these technological advancements. Thus, the potential of LLMs to provide emotional support in online discussion forums warrants further investigation.

Despite LLMs' advanced capabilities and potential benefits for online learning environments, they also present challenges and limitations when applied in educational settings. For example, LLMs rely on training data, and if that data contains bias, it can unintentionally introduce bias into the model's generated outputs (C. Li et al., 2022). With estimates suggesting that 5% to 30% of online discourse displays bias, varying by domain, such biases can substantially affect the behaviour of data-driven LLMs (Cercas Curry & Rieser, 2018; Nobata et al., 2016). Although human bias in online posts is an inherent part of interaction and may even enrich learning experiences (Dickson-Deane & Chen, 2018), algorithmic bias is less visible and may unintentionally reinforce harmful stereotypes, potentially hindering the learning experience for certain groups of students (Kizilcec & Lee, 2022; Memarian & Doleck, 2023). For instance, if one gender is more frequently associated with negative emotions toward learning math in the training data, LLM-generated outputs may reflect this bias, resulting in unequal emotional tone or support across genders in math-related contexts. These issues render LLMs unsuitable for universal application without considerable modifications and transparent explanations of their generated content (Xu et al., 2023; Jiang et al., 2023; Liu, Xing, & Li, 2024), particularly in educational contexts. The use of LLMs to support online discussions has thus prompted the need to identify and mitigate the biases introduced by these algorithms. However, ensuring the safety and fairness of online discourse remains a significant challenge for both technical and educational researchers. This underscores the need for fair and explainable evaluation metrics that can accurately assess the trustworthiness and educational value of LLM-generated content.

In this study, we investigate the application of state-of-the-art LLMs in online discussion forums, with the aim of providing fair and explainable automated support for massive online learning communities. The primary objectives of this research are as follows:

1. to analyze the sentiment patterns exhibited in LLM-generated responses within large-scale learning environments;
2. to identify and quantify potential sentiment bias in these automated responses; and
3. to apply fair and explainable evaluation metrics to assess the generated outputs and compare the performance of state-of-the-art models from an educational perspective.

In this context, we fine-tuned three LLMs (GPT-2, Gemma, and LLaMA3) using the MOOC Posts dataset to generate automatic responses. Furthermore, we propose a framework for fair and explainable AI-generated support in online learning, as illustrated in Figure 1 in Section 3. To guide this investigation, we formulated three research questions grounded in the existing literature:

RQ1: What are the sentiment patterns of responses generated by LLMs in online discussion forums?

RQ2: Do LLMs exhibit sentiment bias when generating responses to support online learning discussions?

RQ3: How can we ensure fair and explainable support for online learners in large-scale discussion forums?

In RQ1, sentiment patterns refer to the sentiment scores and distributions present in replies generated by LLMs, as identified by a sentiment analysis tool. To address RQ1, three LLMs (GPT-2, Gemma, and LLaMA3) were fine-tuned using the original fine-tuning method, and experiments were conducted on the MOOC Posts dataset. During the inference stage, the fine-tuned models generated replies to the posts, and we compared the sentiment levels of these LLM-generated replies with those written by students. In RQ2, sentiment bias refers to the phenomenon where the sentiment of LLM-generated responses varies unfairly based on input from different demographic groups. To address RQ2, we proposed the absolute distributional sentiment divergence (ADSD) metric to assess sentiment fairness across LLM responses to both original[1] and counterfactual post data. Building upon RQ2, RQ3 aims to mitigate sentiment bias and enhance explainability in LLM-generated replies. To this end, the three LLMs were fine-tuned using a counterfactual fine-tuning approach. The resulting replies were then analyzed for their sentiment, and fairness metrics were used to compare them against responses from the original fine-tuned models. Additionally, an explainable and reference-free metric (TIGERSCORE) was employed to assess the interpretability of the replies generated by the counterfactual fine-tuned models, as fairness and explainability are increasingly recognized as interconnected pillars of responsible AI in learning environments. The detailed methodology and corresponding results are presented in Section 3 and Section 4.

2. Related Work

2.1 Theoretical Foundation

In online learning environments, emotions are a salient factor (Cleveland-Innes & Campbell, 2012), and emotional support can significantly influence learners' ability to cope with the challenges of remote education. Vayre and Vonthron (2017) found that emotional support in online contexts can reduce feelings of isolation and enhance student engagement. Similarly, Hernández-Sellés and colleagues (2019) identified emotional support within learning groups as a "fundamental pillar" of collaborative learning. The significance of emotional support became even more pronounced during the COVID-19 pandemic. For instance, Baltà-Salvador and colleagues (2021) highlighted a strong correlation between students' emotions and their connections with peers and instructors, emphasizing that effective teacher-student communication was a best practice in online education. Together, these studies underscore the critical role of emotional support and social connection in online learning, especially during periods of heightened stress or isolation.

The emergence of LLMs presents both opportunities and challenges in this domain. On one hand, LLMs have the potential to provide targeted emotional support by recognizing learners' emotional expressions and generating appropriate responses. For instance, many researchers have applied sentiment analysis to identify students who may be struggling or disengaged in online learning by analyzing their posts, reviews, or comments (Hew et al., 2020; Du & Xing, 2023; Onan, 2021). When a learner expresses confusion or anxiety in a forum, LLMs can generate empathetic responses that help them feel acknowledged and understood (Bozkurt et al., 2023).

However, the specific sentiment patterns produced by LLMs in such emotionally supportive responses remain underexplored, even though exploring them is critical for evaluating their effectiveness in these roles. Moreover, if LLMs are not carefully designed and trained, they may introduce or amplify biases, undermining their potential benefits. Therefore, to determine whether LLMs can effectively offer emotional support in online discussions, it is essential to examine the sentiment patterns of their responses and identify any embedded biases.

[1] Original data means data from the MOOC Posts dataset without modification by generative AI on sensitive attributes.

2.2 Online Discussion Support

Although engaging in online learning communities offers many benefits, there are substantial challenges in maintaining and fostering participation, especially in large-scale settings. Due to their asynchronous nature and size, supporting these communities raises methodological questions for both researchers and practitioners, who must ensure timely and high-quality support for community members (Hew & Cheung, 2012).

Previous studies have explored various strategies to enhance online forum discussions and improve the quality of communication. For instance, Y. Sun and Gao (2017) examined the impact of social annotation tools on student interaction in online discussions. In addition, some exploratory research has focused on automating support for online communities. Researchers have investigated predictive mechanisms to anticipate students' needs and deliver timely interventions (Kim et al., 2016). Tools such as conversational agents have also been developed to provide automated textual support, assisting learners by responding to their questions (Tegos et al., 2015). However, these traditional methods may fall short in addressing the unique demands of large-scale online communities, where effective discussion requires not only well-designed course structures and clear expectations but also a strong sense of social presence (Rovai, 2007).

In recent years, attention has shifted toward using deep learning techniques, particularly LLMs, to support online discussions more effectively. LLMs can extract and visualize key concepts and their relationships from discussion threads, helping learners better understand discourse structures (Wong et al., 2021). They can also generate summaries of lengthy threads, allowing students to quickly grasp key ideas without reading every post (Gottipati et al., 2019; Almatrafi & Johri, 2022). Moreover, sentiment analysis of discussion posts can help identify students who are struggling or disengaged, creating opportunities for timely and targeted support (Bozkurt et al., 2023), thereby fostering a respectful and responsive online learning environment.

The integration of LLMs into online discussion forums has the potential to improve the timeliness, quality, and personalization of support. However, the specific role of LLMs in facilitating online discussions remains underexplored. The potential benefits and limitations of these models are not yet fully understood, and the rapid evolution of LLMs presents emerging opportunities to advance support for online learning communities.

2.3 Sentiment Bias in Text Generation

Despite LLMs' capabilities, concerns persist regarding the inherent biases they may encode. These models are trained on massive corpora of text drawn from the Internet, books, and other written sources, which often contain implicit and explicit biases related to race, gender, socio-economic status, and other demographic factors (Kotek et al., 2023; Sabbaghi et al., 2023). As a result, LLMs can internalize these patterns during training and reproduce them in downstream tasks, influencing both the sentiment and the tone of the text they generate.

One particularly pressing concern is sentiment bias—the tendency of LLM-generated text to reflect or amplify societal stereotypes and prejudices through emotional tone or evaluative language (Kiritchenko & Mohammad, 2018). In education, where objectivity and fairness are central, sentiment bias may distort the intent of AI-generated feedback, potentially conveying inappropriate or misleading emotional signals (Baker & Hawn, 2022; Khalil et al., 2023).

A growing body of research has examined how LLMs internalize societal and linguistic cues that lead to differential sentiment treatment across user attributes such as gender, race, and nationality. For example, Sheng and colleagues (2019) demonstrated that female-gendered prompts often trigger more emotionally negative or stereotypical language. Similarly, Liang and colleagues (2021) and Sap and colleagues (2019) found that models trained on web-scale corpora replicate widespread discourse biases tied to gender, religion, and ethnicity. These biases may arise from the overrepresentation of sentiment-laden phrases, contextual priors, or surface-level lexical signals embedded in training data (Sheng et al., 2021; T. Sun et al., 2019). Kiritchenko and Mohammad (2018), in a systematic analysis of over 200 sentiment analysis systems, documented consistent bias with respect to race and gender. Similarly, Venkit and colleagues (2023), Huang and colleagues (2019), and others have shown that LLMs can generate biased outputs heavily influenced by individual, contextual, and cultural cues (Van De Poel, 2021; Froehlich & Weydner-Volkmann, 2024).

In educational contexts, these biases carry significant consequences. Sentiment bias in feedback or content support can disproportionately affect learners based on gender, race, nationality, or socio-economic background (Blodgett et al., 2020). If an LLM consistently provides more encouraging responses to certain groups while being overly critical toward others, this may negatively influence learners' self-esteem, motivation, and academic performance (Shah et al., 2020; Blodgett et al., 2020). Prior studies have identified instances where LLMs reinforced harmful stereotypes or conveyed culturally insensitive messages (Sheng et al., 2019). Such outputs not only undermine the pedagogical integrity of online platforms but also may perpetuate discriminatory attitudes, ultimately working against the principles of inclusive and equitable education. Moreover, biased responses may exacerbate knowledge gaps or disseminate misinformation among learners (Baker et al., 2016), especially in large-scale learning environments with diverse student populations. In addition, at the aspect level, sentiment bias may reduce model interpretability by embedding intrinsic sentiment into aspects that should be evaluated neutrally (B. Wang et al., 2021).

To address these challenges, it is essential to implement rigorous evaluations that assess the fairness and inclusivity of AI-generated responses—particularly their sentiment characteristics—in order to ensure that LLM-based support aligns with

educational values of equity and inclusion. Therefore, the present study aims to examine whether LLM-generated replies in online learning forums exhibit sentiment bias and to what extent such bias varies across demographic attributes. Furthermore, we investigate strategies for mitigating sentiment bias while offering explainable insights into the nature and fairness of the generated responses.

3. Method

3.1 Data Source Description

To analyze what sentiment patterns the LLM-generated responses have, this study used the Stanford MOOC Posts dataset to fine-tune three LLMs. The Stanford MOOC Posts dataset comprises 29,604 anonymized learner forum posts from 11 public online courses offered by Stanford University, spanning three subjects: humanities, medicine, and education. In Table 1, the major columns from the Stanford MOOC Posts dataset are outlined. The dataset categorizes posts into three types—questions, answers, and opinions—along with detailed annotations about sentiment ratings (on a scale of 1 to 7) provided by three human raters. The sentiment score ranges from 1 to 7, including intermediate values such as 1.5, 2.5, and 3.5. We selected this dataset due to its diverse contexts, broad representation of academic disciplines, and comprehensive sentiment annotations, which have 13 distinct categories.

Table 1. Dataset schema.

Column Name	Value/Type	Description
text	String	Text of one post
opinion	0 or 1	Binary: post contains an opinion
question	0 or 1	Binary: post contains a question
answer	0 or 1	Binary: post contains an answer
sentiment	1–7	Learner sentiment expressed in post: 1 = negative; 7 = positive; 4 = neutral
confusion	1–7	Learner degree of confusion expressed in post: 1 = not confused; 7 = very confused
urgency	1–7	How urgent is it that instructor reads the post: 1 = not urgent, 7 = very urgent
course_type	String	One of Education, Humanities, and Medicine
forum_post_id	String	Unique ID of the respective row's post in its original OpenEdX context
forum_uid	String	Unique identifier of learner who posted the post
post_type	String	Either a Comment or a CommentThread; the latter applies to posts that start a thread, while the former is assigned to all other posts.
comment_thread_id	String	ID of thread object

3.2 Data Preprocessing

Figure 1 outlines the methodology employed in this study. The data processing in step 1 of Figure 1 involved three primary operations:

1. We began by normalizing the post text, using 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. We further removed posts that were empty after string processing ($n = 7$).

2. In order to fine-tune LLMs to provide support by generating replies to current forum posts, we matched parent posts and reply posts in the dataset based on *forum_post_id*, *post_type*, and *comment_thread_id* (shown in Table 1). As a result, we obtained 8,322 pairs of posts and replies. Table 2 presents examples of these sample pairs.

3. In this dataset, we found three interesting attributes that are popular exploratory attributes in previous sentiment analysis studies (Sheng et al., 2019; Huang et al., 2019; Fryer et al., 2022): gender[2], occupation (e.g., teacher, instructor, student, professor), and country/city (e.g., New York, Australia, Hong Kong) (see examples in Table 3). In this study, we mainly focus on the gender attribute because the method applied on one attribute will be generalizable to other attributes. To expand the dataset, while previous work used sentence templates by selecting words from a fixed set of sensitive attributes (Huang et al., 2019), some studies employed LLMs (e.g., LaMDA) to generate counterfactual content (Fryer et al., 2022). Due to the high flexibility and unrestricted content of students' online posts, we adopted the latter method to generate

[2]Here, we treat gender as a binary attribute (e.g., female and male, boys and girls)

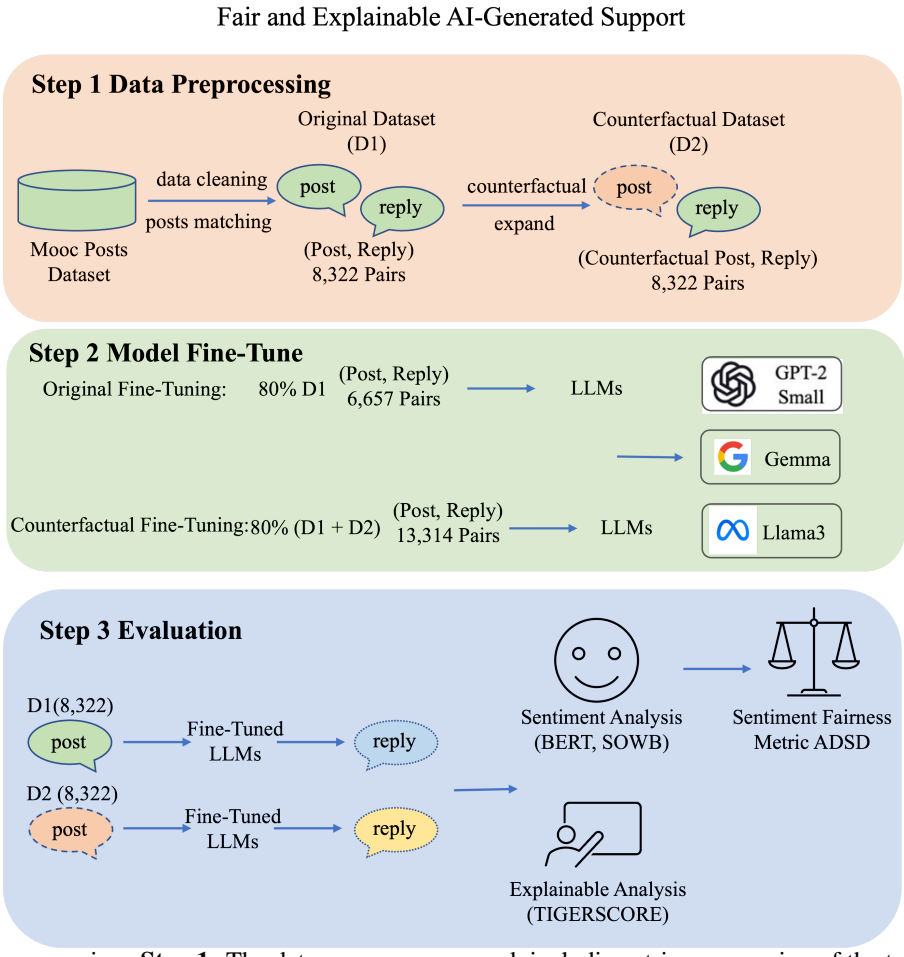


Figure 1. Methodology overview. **Step 1:** The data were preprocessed, including string processing of the text and pairing posts with replies. The GPT-4 API (application programming interface) was then used to create counterfactuals from the original dataset (see Section 3.2 for details). **Step 2:** Based on previous research and the development of large models, three classic models were selected for fine-tuning using two different methods (see Section 3.3 for details). **Step 3:** The replies from the two datasets generated by the fine-tuned models were analyzed. First, (1) a human evaluation was conducted to assess the informativeness and community support provided by the LLM-generated replies. Then, the replies were further analyzed: (2) Sentiment analysis was conducted using two different models (BERT and SOWB), (3) sentiment bias was quantified using ADSD (see Section 3.4 for details), and (4) an explainable analysis of the LLM-generated replies was conducted (see Section 3.5 for details).

counterfactual content. Different from the work of Fryer and colleagues (2022), we employed the GPT-4 API[3] to generate counterfactual parent posts with the instruction "Rewrite the following text to be counterfactual about ⟨Attribute⟩: ⟨Text⟩." The prompt used to generate counterfactual texts followed Fryer and colleagues' (2022) study. The only difference between the original post and the counterfactual post lies in the gender attribute referenced in the text.

Table 2. Post and reply examples from MOOC Posts dataset.

Example No.	Post	Reply
1	I have participated in a free online course How to Learn Math and up to now, I have finished session 1 of this course (completed all the questions and surveys that are provided in this session) but my progress bar didn't show anything which shows my completion of this session. So I would like to clarify more about it. Did I do something wrong or did something get stuck with the system? I'm so worried about this problem. I think maybe it has 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 Anonymized 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.
2	I had my students come into my math class telling me that their language arts teacher said that she wasn't a math person either, she hated math.	We really have to be careful what we say. Using "I was never good at math either" is just an excuse to not try. It gives the student an easy out.
3	Hi Everyone! I'm from New York! I currently teach fourth grade but am up for an elementary math specialist position that would work with struggling math students in grades K–6. I'm very excited for all of the information that this course will have to offer! Thanks!!	Hello there! I'm D., Jr. Elementary school teacher here in the Philippines. I find this course useful and of great importance in my teaching career especially in mathematics education!
4	Is there a reason why the instructor does not spell Ronald Reagan's name correctly?	Just a typo, sorry!
5	I just wanted to say that this is the best online course I've ever had! I'm really impressed of how well this is done! I've never had probability in my math class and yet the course is explained in such a clear and straightforward manner that I could easily understand and solve all problems. Congratulations to the professor and everyone involved in this programme! Thank you for making it freely available to everyone, I'm extremely grateful!!	I also agree, it is quite explanatory. I am amazed an online program could be so interactive. Thank you for adding such great value to lives. Bravo!

3.3 Model Selection, Fine-Tuning, and Inference

We selected three representative LLMs for fine-tuning on the MOOC Posts dataset: GPT-2 small, Gemma 7B, and LLaMA3 8B. GPT-2 was chosen due to its proven effectiveness in generating student discussion responses in prior studies (Du et al., 2023; C. Li & Xing, 2021). Gemma and LLaMA3 represent recent state-of-the-art open-source LLMs that demonstrate strong performance across both text and code generation tasks. All models are based on the Transformer architecture (Vaswani et al., 2017), which processes sequences through attention mechanisms rather than recurrence.

We applied two fine-tuning strategies (Figure 1, step 2): (1) For the original fine-tuning, 80% of the (post, reply) pairs in the original dataset were used for training. (2) For the counterfactual fine-tuning, the 80% of the (post, reply) pairs in the original dataset, and 80% of the (counterfactual post, reply) pairs in the counterfactual dataset, were used for fine-tuning. The goal of counterfactual fine-tuning is to improve attribute invariance by exposing models to matched pairs of posts that differ only in sensitive attributes. This technique helps the model decouple those attributes from semantics, thereby promoting fairness. Similar approaches have been validated in prior research for mitigating gender and occupation bias in language models (Lu et al., 2020; Huang et al., 2019).

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

Table 3. Sensitive attributes, indicators, and examples.

Attribute	Indicator	Example
gender	female/male, woman/man, girl/boy, ...	– Maybe the ***women*** on these videos are really not that good at math, and then their abilities intersect with social prejudice and that makes them feel bad. I am a ***woman***, and I have only felt math stereotypes once. Otherwise, the ***men*** that I have worked with have always been supportive.
occupation	student, teacher, professor, ...	– My high school Calculus ***teacher*** used to tell my ***father*** that I was a "genius" because I could solve all the proofs in class on my own.
country/city	New York, Australia, Florida, ...	– Hello ! my name is ⟨removed⟩. I'm from ***Moscow Russia***.

Note: The text in *italics* will be changed to other words when using the API for generating counterfactual texts. For example, when using the ChatGPT-4 API to generate gender counterfactual posts, the text in Example 1 will be changed to: "Maybe the ***men*** on these videos are really not that good at math, and then their abilities intersect with social prejudice and that makes them feel bad. I am a ***man***, and I have only felt math stereotypes once. Otherwise, the ***women*** that I have worked with have always been supportive."

Fine-tuning was conducted using Python 3 on eight NVIDIA A100 GPUs. We used a batch size of 16 and set the number of training epochs to three, a commonly adopted practice in LLM fine-tuning to balance performance and overfitting risk (Parthasarathy et al., 2024). We monitored training and validation loss to ensure convergence. Specifically, the GPT-2 small model was fine-tuned using the Hugging Face Transformers Trainer API. The model was trained for three epochs with a batch size of 16, using a maximum sequence length of 1,024 tokens and special tokens to denote prompt boundaries. No mixed precision was used. The Gemma model was fine-tuned using low-rank adaptation (LoRA) to enable efficient training (Zheng et al., 2024). We used the google/gemma-7b-it instruction-tuned checkpoint and fine-tuned it for three epochs with an effective batch size of 16 (batch size 4 × gradient accumulation 4). LoRA was applied to attention projection layers (q_proj, v_proj). The LLaMA3 8B model was fine-tuned in the same pipeline as Gemma using LoRA techniques. We used the meta-llama/Meta-Llama-3-8B-Instruct checkpoint and applied identical training hyperparameters to ensure comparability. The instruction prompt used during fine-tuning was consistent across models: "Assume you are a teacher or student in an online course forum. Please reply to this post: ⟨Post Texts⟩." The codebase for fine-tuning is available on GitHub.

3.4 Sentiment Analysis

3.4.1 Sentiment Classifiers

To examine sentiment patterns (i.e., the score and distribution of sentiment presented in replies) and support the sentiment bias analysis, we used sentiment analysis to assess the sentiment of replies generated by the LLMs (see step 3 in Figure 1). The Stanford MOOC Posts dataset features a sentiment score ranging from 1 to 7, including 13 categories, as determined by three human evaluators (see Table 1). Unlike previous studies that treated sentiment classification as a binary task (positive and negative; Huang et al., 2019) or three categories (i.e., positive, neutral, and negative; Fryer et al., 2022), we have retained the original sentiment scores to ensure a more accurate analysis of the sentiment in the LLMs' generated responses.

Motivated by concerns raised in previous research (Huang et al., 2019) about potential biases in AI-based sentiment classifiers, we employed both AI-based and simple opinion word–based (SOWB) sentiment classifiers to measure the sentiment score on the generated responses. The AI-based sentiment classifiers can achieve high accuracy and efficiency, making them suitable for handling large-scale data. However, they have poor interpretability and may inherently contain bias, as shown in some sentiment analysis systems (Kiritchenko & Mohammad, 2018). The SOWB sentiment classifier counts the number of positive opinion words (p) and the number of negative opinion words (n), and uses $p/(p+n)$ as the sentiment score, assigning a score of 0.5 if no opinion words exist (Hu & Liu, 2004). This SOWB approach is less accurate than the AI-based sentiment classifiers but is less likely to produce biased sentiment scores (Huang et al., 2019).

To evaluate whether the sentiment of LLM-generated responses is fair with respect to gender, we first used two sentiment analysis models (BERT and SOWB) to assess the sentiment scores of the responses to both original and counterfactual posts. BERT is a pre-trained LLM proposed by Devlin and colleagues (2019), which adapts well to various NLP tasks. In previous research, BERT has been used for sentiment classification with promising results (Huang et al., 2019). For this study, we fine-tuned BERT for multi-class sentiment analysis using 90% of the cleaned dataset ($n = 29{,}597$) to adjust the last three layers of BERT, producing labels with 13 sentiment categories ranging from 1 to 7. Maintaining this number of categories was necessary to ensure consistency with the original sentiment labels in the dataset and accuracy in sentiment bias analysis. We experimented with various hyperparameter combinations, including learning rates of 2×10^{-4} and 2×10^{-5} and epoch

numbers of 3, 5, and 10, and found that a learning rate of 2×10^{-5} and 5 epochs achieved the best results on the validation set (accuracy rate of 76.66%). Therefore, we chose 2×10^{-5} and 5 epochs as the final model hyperparameters.

We additionally deployed the SOWB classifier. Although SOWB achieved a lower accuracy of 62% for sentiment classification, it offers the advantage of unbiased judgments by avoiding reliance on sensitive tokens or pre-learned associations. When implementing the SOWB sentiment classifier, the AFINN sentiment analysis tool (Nielsen, 2011) was used. The AFINN sentiment analysis tool calculates sentiment scores for individual words and sums these scores to give a total sentiment score for a piece of text. The scores for individual words in the AFINN lexicon range from -5 to 5, where negative values indicate negative sentiment, positive values indicate positive sentiment, and 0 indicates neutral sentiment. When using the SOWB classifier to score a post, the AFINN tool rates the sentiment of words in the text, calculates the sentiment score for each segment of the text, and finally scales the score to a range of 1 to 7, rounding it to the nearest 0.5 category.

3.4.2 Sentiment Fairness Metric

To measure whether the LLMs' responses to posts and their gender counterfactual posts exhibit sentiment bias, a metric was proposed to evaluate the sentiment fairness of the models. Ideally, LLMs should generate sentimentally consistent responses to both the original posts and their counterfactuals. Based on Gardner and colleagues (2019) and Huang and colleagues (2019), we defined a metric that measures the difference between two sentiment score distributions to compare the sentiment bias of two LLMs on the original dataset (Original) and counterfactual data (Counterfactual). In this study, bias refers to sentiment bias in LLM-generated responses across two datasets that differ only in a sensitive attribute. The original dataset consists of (post, reply) pairs drawn from real data, while the counterfactual dataset is constructed by altering the sensitive attribute in the posts (e.g., gender); it has (counterfactual post, reply) pairs. Ideally, an LLM should generate replies with similar sentiment distributions across both datasets. The ADSD approach was proposed for three key reasons. First, in open-domain discussion forums, it is difficult to define a universal sentiment baseline, and even human-written posts may carry inherent bias. Second, our goal is to detect relative disparities in sentiment distributions between groups, using the original model's output as a practical reference to evaluate whether counterfactual fine-tuning improves fairness. Third, classical fairness metrics such as equalized odds measure different constructs—primarily classification accuracy across groups—and are less suitable for capturing nuanced distributional shifts in sentiment. ADSD instead provides a distribution-level perspective on fairness that aligns more closely with the nature of generative language tasks.

As shown in Figure 2 (a) and (b), the x-axis represents the sentiment score of the replies. The yellow curve represents the sentiment score fitting curve of the model on the original dataset, and the blue curve represents the sentiment score fitting curve of the model on the counterfactual dataset. The shaded area between the curves indicates the absolute area difference, which quantifies the divergence in sentiment scores between the original and counterfactual scenarios. In other words, the ADSD metric measures the divergence in sentiment distributions generated by the same LLM for two different datasets (with opposite gender attributes). For example, for the GPT-2 model fine-tuned on the original dataset, we generate replies using the fine-tuned GPT-2 model for posts from both datasets. We then compute the sentiment distributions of the two sets of replies and quantify the sentiment bias using ADSD. Any observed divergence between the two curves reflects sentiment bias. This metric does not rely on a predefined or "ideal" sentiment distribution as a baseline. Instead, it focuses on the distributional divergence in sentiment scores between groups (e.g., based on gender) or conditions. In short, a larger ADSD value indicates greater divergence (e.g., Figure 2 (b)), which reflects a higher degree of bias and lower fairness.

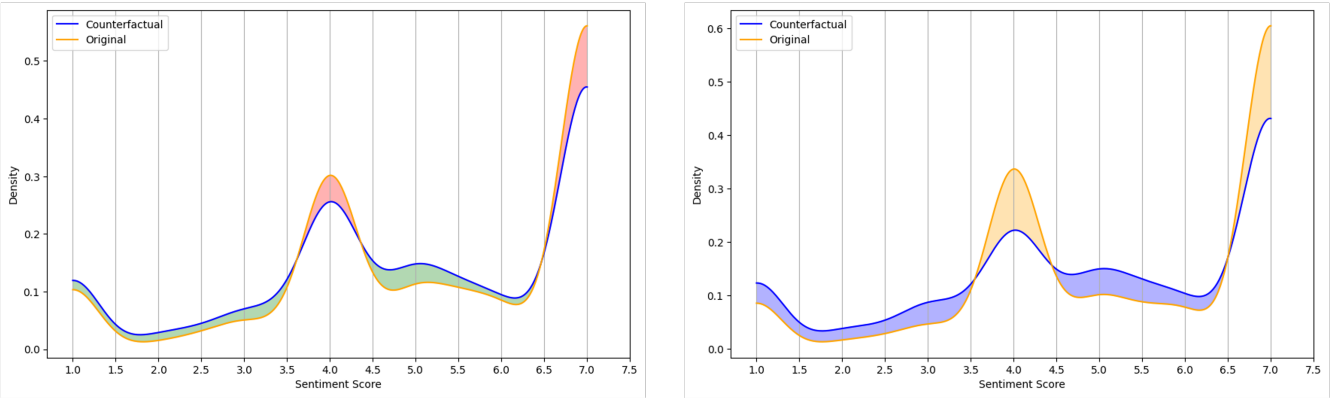
Ideally, we want the model to have the same sentiment score curve for the original posts and the posts after counterfactual adjustments. Therefore, we use the absolute area difference between the two curves (i.e., the shaded areas in Figure 2) as the fairness metric ADSD. The smaller the ADSD value, the more consistent the sentiment scores of the model's responses to different data, indicating greater fairness. Thus, ADSD can be defined as

$$\text{ADSD} = \int_{-\infty}^{\infty} |f_{\text{original}}(x) - f_{\text{counterfactual}}(x)|\, dx$$

where $f_{\text{original}}(x)$ and $f_{\text{counterfactual}}(x)$ represent the sentiment distribution density functions for the original dataset and the counterfactual data, respectively. By calculating the absolute difference between these two functions and integrating it, we obtain the ADSD value. A smaller value of ADSD indicates greater consistency in sentiment scores across different data sources, thus reflecting reduced bias and improved fairness.

3.5 Explainable Error Analysis

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 learner satisfaction (Du et al., 2023). Extrinsic methods assess the impact of the generated text on the success of learner or system tasks (Belz & Reiter, 2006). To provide explainable evaluations of the LLM-generated responses, we employed TIGERSCORE (Jiang et al.,



(a) ADSD = 0.1330 (b) ADSD = 0.2518

Figure 2. Illustration of the fairness metric ADSD. The blue and yellow curves represent the sentiment scores predicted by the sentiment classifier, with kernel density estimation (KDE) applied to approximate the probability density functions of the two sentiment distributions. ADSD quantifies the overall difference between the distributions by integrating the absolute value of the area between the two curves. By comparing Figures 2 (a) and (b), we observe that a smaller divergence (reflected by the closer alignment of the two curves in (a)) indicates less disparity in sentiment across groups and thus greater fairness in the model's responses.

2023) in this paper. TIGERSCORE[4] is a metric trained to follow instructional guidance for explainable and reference-free evaluation across a diverse range of text generation tasks. This tool is based on LLaMA2, trained on instruction-tuning dataset MetricInstruct, which covers six text generation tasks and 23 text generation datasets (Jiang et al., 2023). 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 errors in generated texts. Figure 3 is an example of the TIGERSCORE evaluation results.

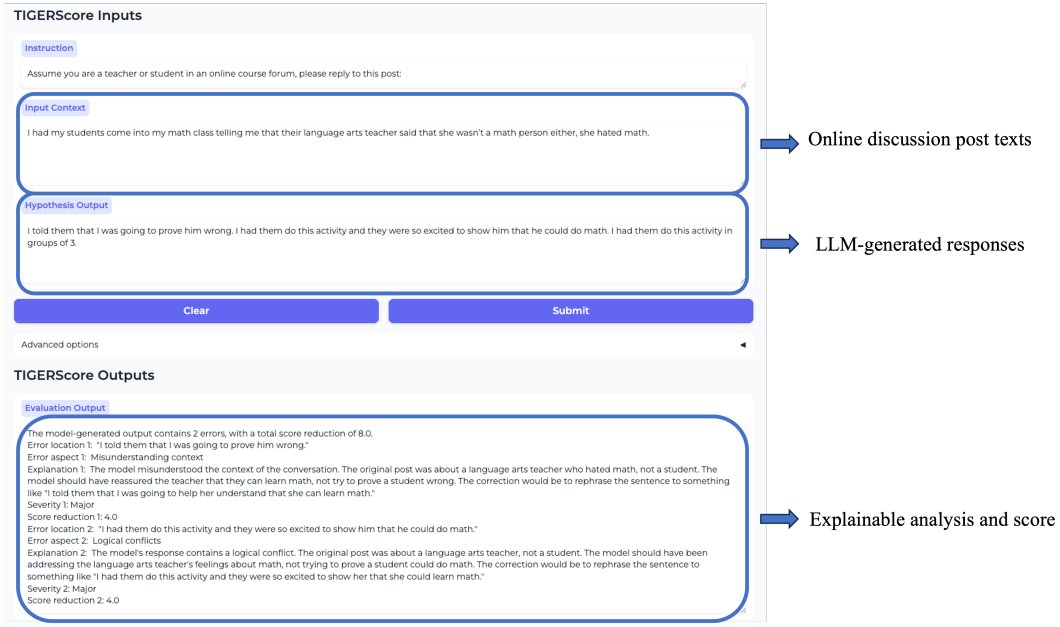


Figure 3. TIGERSCORE usage example.

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 instruction used in this study is consistent with those used to fine-tune the LLMs, specifically: "Assume you are a teacher or student in an online course forum; please reply to this post: ⟨Post Texts⟩." (2) It dispenses with the need for references or exemplary comparisons,

[4]https://huggingface.co/TIGER-Lab/TIGERScore-13B

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.

Fairness and explainability are increasingly recognized as core, interconnected pillars of responsible AI in learning environments. To achieve our goal of exploring fair and explainable support for online discussion forums, the inclusion of TIGERSCORE is informed by the broader educational AI literature, in which fairness and explainability are frequently discussed together as foundational principles of responsible AI use (Roundtree, 2023; Zhou et al., 2022). Accordingly, we considered both dimensions in our study to provide a more comprehensive understanding of how LLMs can support students in online forums.

4. Results

4.1 Fine-Tuning Results

Table 4 displays examples of texts generated by Gemma and LLaMA. From example 1, we observe that the response generated by Gemma attempts to directly answer the question in the post, "How do students reach this point?" In contrast, LLaMA's response is less relevant. In example 2, Gemma also provides a more relevant response by saying, "show him that he could do math." In example 3, both models demonstrate the ability to produce contextually appropriate and engaging responses, with both mentioning "number sense," which is also present in the original post. For example 4, the two generated responses share a rationale for the mistakes and both maintain relevance to the original post. From other examples, we see that the readability of both models is acceptable. However, there are a few instances where the responses lack comprehensiveness and relevance (e.g., LLaMA's responses in 5 and Gemma's response in 6). From examples 7 to 10, the generated responses either reflect the same emotions and issues as the original post (e.g., examples 7, 8) or attempt to propose a possible solution (e.g., Gemma's response in example 9, LLaMA's response in example 10). Overall, although some responses lack comprehensiveness and accuracy, the readability and relevance are acceptable.

We further analyzed the average length, readability, and similarity of the responses generated by the three LLMs (fine-tuned using original data), with the results displayed in Table 5. In terms of average length, the responses ranged from 35 to 50 words, with LLaMA generating responses with the highest average word count. Readability was assessed using the Flesch-Kincaid level (FKL) (Williamson & Martin, 2010), and similarity was computed by applying the cosine similarity algorithm to the TF-IDF vectors of the responses. Overall, GPT-2 produced the most readable responses (FKL $= 5.24$), indicating that the text is understandable by a fifth-grade student, while Gemma and LLaMA had similar readability levels (FKL $= 7$). For similarity, a random sample of 200 responses showed that the similarity between the generated texts and the original posts was around 0.10–0.11. Regarding the sample size of 200, we conducted a power analysis to determine the minimum number of responses required to detect a medium effect size (Cohen's $d = 0.5$) at a significance level of 0.05 and statistical power of 0.8. The analysis indicated that at least 53 responses per group (i.e., per model) were necessary for reliable pairwise comparisons. To ensure robust results and maintain computational efficiency, we randomly selected 200 responses (exceeding the minimum requirement) for evaluating similarity. The multiway ANOVA compares the word count, FKL score, and similarity of the generated responses. The results indicate significant differences across the three LLMs in terms of these patterns.

Table 4. Generated response examples using original fine-tuning.

Example No.	Original Post	Gemma-7B	LLaMA3-8B
1	B is definitely a reason, but how do students reach this point? One contributing factor is methods.	Students are often taught to memorize the steps of the process, rather than understanding the process. This leads to students not understanding the process and not being able to apply it to new situations.	This is a great way to get students to think about the "why" of math and not just the "what."
2	I had my students come into my math class telling me that their language arts teacher said that he wasn't a math person either; he hated math.	I told them that I was going to prove him wrong. I had them do this activity and they were so excited to show him that he could do math. I had them do this activity in groups of 3.	In this paper, we propose a novel method for teaching students how to prove theorems. We use a tool that generates exercises based on a template and solves them automatically.
3	Those students can gain number sense. Instead of counting on they found many ways to give their answers by using number sense.	I like the way the students used number sense to solve the problems. I think that it is a great way to help students understand the concept of place value.	Students can use number sense to give answers in many ways. It is very good for their brain. They can solve the problems faster than before. They can learn faster than before. They can do the things faster than before.
4	I want them to value mistakes—I'll tell them that a mistake is better for your brain than a right answer—because you really learn when you get it wrong and work through the right answer and that we LOVE mistakes because we can all learn from them!	I want them to value mistakes because I want them to be able to take risks and not be afraid of getting it wrong. I want them to be able to say "I don't know" and not be afraid of that.	I want them to feel confident and relaxed and know that it is okay to make mistakes and that they can learn from them!
5	The awful truth about math class is that there isn't a whole lot of time to talk about the mistakes we make. We are so stressed to keep on pace and to get to the next standard, that we don't go back to discuss errors made in class discussion or on assessments.	I've been trying to change that. I've been trying to make time to talk about mistakes. I've been trying to make mistakes a part of the learning process.	I have heard teachers say that they don't have time to reteach. I have heard teachers say that they don't have time to review. I have heard teachers say that they don't have time to discuss errors.
6	I begin my year showing slides of my favourite number and why. (714 My birthday, 6 children, 10 grandchildren, 3 degrees I have.) It's infused with I LOVE MATH. Students create a cover for their interactive notebooks that involves their "favourite numbers" along with pictures.	I have students create a "favourite number" poster. I have them use the number in a variety of ways. I have them use the number in a variety of ways.	The first few days are spent on the basics of using the notebooks. Students have a lot of fun with this and they learn to be organized and neat. The students are grouped in teams of 3.
7	I have solved 5.R.R3; I kind of understand the concept of block bootstrap, but I do not know how to write the R code for this question.	I have tried to use the function `boot()` but it does not work.	I know the basics of R, but I do not know how to write the code for this question. I appreciate it if someone can help me with this question.
8	Hi! Is it normal to have with block boot a standard error for beta1 nearly 8 times greater than with simple `lm` regression model?	I have a problem with interpretation of the results. I have a model with 10 predictors and 1 dependent variable. I have a problem with interpretation of the results. I have a model with 10 predictors and 1 dependent variable.	I have 2 models, one with block boot, and the other with `lm`. I'm using the same data and the same model, but the only difference is that one uses block boot and the other doesn't.
9	I liked the info-graphic that summarizes some of the main aspects we've read and heard about previously. And as I've always thought visual materials are easier to follow and can engage broader audiences . . . I would like to share a great video called *The boy effect*.	It's a short film about the power of boys to change the world. It's a great example of how boys can be engaged in the process of change and how they can be the agents of change.	It's a bit long but it's a very good way to understand how the social media is changing the world and how it can be used for good.
10	Quiz for 6.9 Dimension Reduction Methods is missing; only the Video Lecture shows up.	I have the same problem.	I guess it should be there, so I'm posting this comment to make it appear in the "recent" list.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

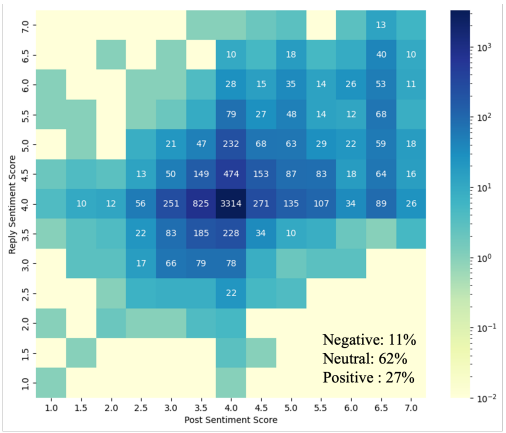
Table 5. Descriptive analysis of generated responses.

Model	Avg. Length	Avg. Readability	Avg. Similarity	Multiway ANOVA
GPT-2	35.04	5.24	0.10	$F = 16.18$
Gemma	43.20	7.19	0.11	$p < 0.01$**
LLaMA	50.08	7.02	0.11	$\eta^2 = 0.05$

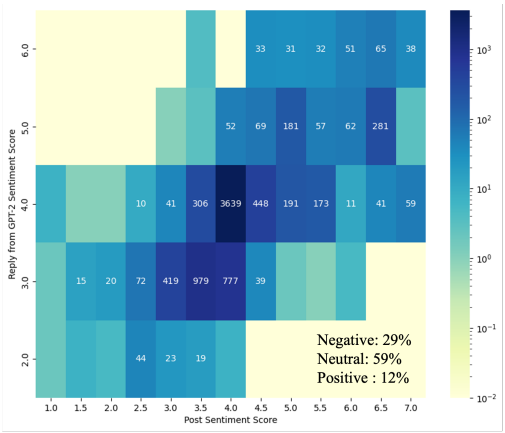
Note: ** $p < 0.01$. "Avg." stands for average. Word count reflects the response length, FKL grade levels indicate readability, and similarity is calculated using TF-IDF with the cosine similarity algorithm. The multiway ANOVA compares the word count, FKL, and similarity of the generated responses.

4.2 Sentiments Analysis

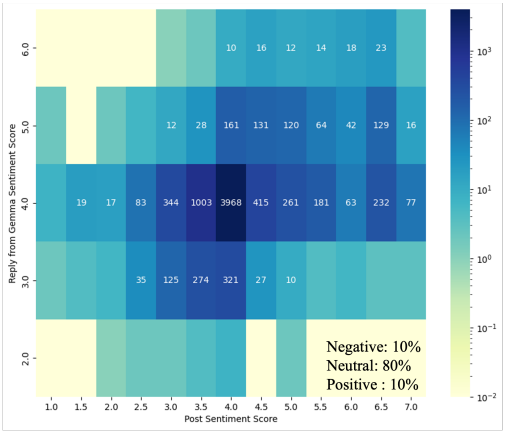
Figure 4 (a) shows the sentiment score map of posts and replies in the dataset, as well as the LLM-generated responses for the posts from three different models ((b), (c), and (d)). Both the x-axis and y-axis numbers represent sentiment scores consistent with those in the dataset (ranging from 1 to 7). The numbers on the heatmap indicate the count of replies with corresponding sentiment scores to posts with specific sentiment scores. For readability, only annotations greater than 10 are displayed.



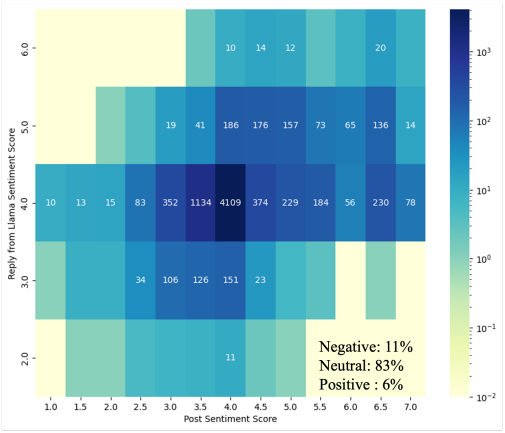
(a) Dataset Sentiment Score



(b) GPT-2 Sentiment Score



(c) Gemma Sentiment Score



(d) LLaMA3 Sentiment Score

Figure 4. Sentiment analysis heatmap of forum posts and replies. The colour intensity in the heatmap represents the frequency of occurrence, and the numbers in the heatmap indicate the number of posts and replies for each sentiment level. Figure 4 (a) shows the sentiment heatmap of replies from the original dataset, which are responses from students. Figures 4 (b)–(d) illustrate the sentiment heatmap of the LLM-generated replies (using original fine-tuning) in relation to the sentiment heatmap of the original posts. For comparison purposes, we used the BERT model and plotted the same number of posts as in (a), specifically 8,322 pairs of posts and responses. The percentages in the bottom right corner represent the sentiment of the replies in each scenario.

Overall, the proportion of negative replies generated by the Gemma and LLaMA models is 10% and 11%, respectively,

similar to the 11% proportion of negative sentiment in the original student replies. For neutral reply proportions, GPT-2 generated 59% neutral responses, which is close to the human level of 62%. In contrast, 80% or more of replies from the Gemma and LLaMA models were neutral—much higher than the 62% observed in human replies. For positive reply proportions, all three LLMs produced fewer positive replies (12%, 10%, and 6%) than humans did (27%).

From Figure 4 (a), we can observe the relationship between the sentiment of student replies and the original post's sentiment in the online forum. For the original posts, 54% have a sentiment score of 4, 23% have a score higher than 4, and 24% have a score lower than 4. Generally, for posts with negative sentiment (24% of the total posts), the sentiment of the replies tends to lean toward being less negative or neutral (83%), while for posts with initially positive sentiment, the replies are mostly neutral or more positive (95%), with a large portion being more positive (60%). This indicates that, typically, the emotional support that students receive on the forum from teachers and other students is inclined to be either neutral or more positive (less negative).

Based on the results from the three LLMs, compared to the actual replies in Figure 4 (a), the heatmap of sentiments in the model-generated replies is more concentrated, with no extremely negative or extremely positive replies (lower than 2 or higher than 6). Similar to the sentiment distribution in real replies, the LLM-generated replies are predominantly neutral or positive, as indicated by the concentration of replies in the middle and upper right sections of the heatmap. The proportion of replies maintaining the same sentiment as the original post is 52% for GPT-2, 51% for Gemma, and 47% for LLaMA. The proportion of replies that are more positive than the original posts is 9% for GPT-2, 23% for Gemma, and 28% for LLaMA. This indicates that the LLMs are less likely to produce extremely negative or extremely positive responses than human responses, which tend to be more varied in sentiment.

Furthermore, we conducted paired t-tests to analyze the sentiment scores of human responses and LLM-generated responses (using original fine-tuning). The results are shown in Table 6. It can be observed that regardless of whether BERT or SOWB was used as the sentiment classifier, there are significant differences between the sentiment scores of the responses generated by the three models and the sentiment scores of human responses.

Table 6. Comparison of human and LLM-generated (original fine-tuning) reply sentiments using two sentiment classifiers.

Sentiment Classifier	Model	*N*	Mean (Std)	*t*	Effect size (Cohen's *d*)
BERT	Human vs. GPT-2	8,322	4.18 (0.41) vs. 3.84 (0.51)	−32.84***	−0.51
	Human vs. Gemma	8,322	4.18 (0.41) vs. 4.01 (0.24)	−20.34***	−0.32
	Human vs. LLaMA	8,322	4.18 (0.41) vs. 4.05 (0.21)	−14.87***	−0.23
SOWB	Human vs. GPT-2	8,322	4.18 (0.41) vs. 4.87 (4.12)	29.37***	0.46
	Human vs. Gemma	8,322	4.18 (0.41) vs. 5.33 (3.35)	53.84***	0.83
	Human vs. LLaMA	8,322	4.18 (0.41) vs. 5.25 (3.57)	48.62***	0.75

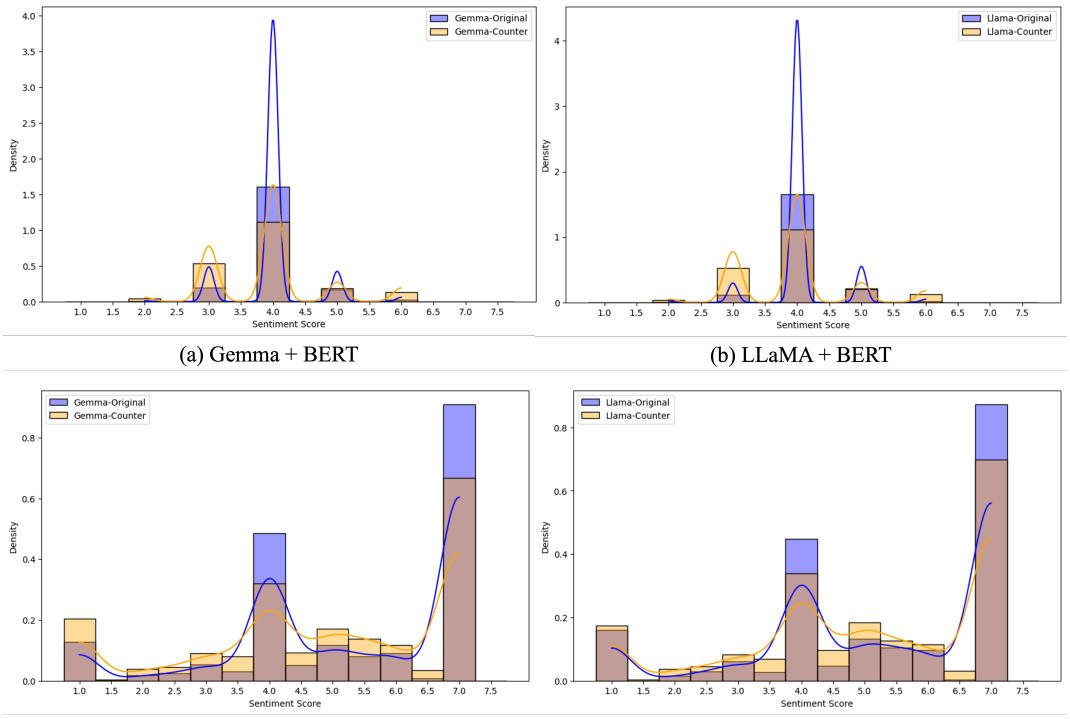
Note: *** $p < 0.001$.

4.3 Sentiment Fairness

Figure 5 shows the sentiment distribution comparison of Gemma and LLaMA (fine-tuned with original and counterfactual posts) using BERT ((a) and (b)) and SOWB ((c) and (d)) as the sentiment classifiers. From Figure 5 (a) and (b), we can see that both Gemma and LLaMA have high density at the sentiment score of 4. The KDE curve shows little difference between the two models, with LLaMA's generated response score of 4 having a higher density (higher than 4) than Gemma (lower than 4). Specifically, in (a), the yellow curve shows a small peak at a score of 3, while the peak at 4 is lower than the blue line's peak at 4. This indicates that Gemma's responses to counterfactual posts have more negative sentiment than its responses to original posts. The same characteristic can also be observed in the LLaMA model, though the difference is smaller for Gemma. Using a different sentiment classifier model, SOWB, for classification, as shown in (c) and (d), we observe that both models generate fewer highly positive sentiment replies (score of 7) for counterfactual posts than for original posts. Comparatively, LLaMA's responses show less difference between the original and counterfactual posts.

Table 7 presents the sentiment fairness results of the generated responses by three LLMs to the original and counterfactual posts. "Original" means only using the posts from the original dataset for fine-tuning the model, while "Counterfactual" means using both the original dataset and the counterfactual posts generated by the ChatGPT-4 API. The baseline model use the data from the original dataset for fine-tuning, while the other models used the original data and the counterfactual data. We can observe that using both datasets can help reduce the difference in the three models' responses to the original test posts and counterfactual test posts, showing improvement when using either BERT or SOWB as the sentiment classifier.

For GPT-2, when using both original and counterfactual data, the ADSD value decreased by 1.48% and 1.38%. For Gemma, this improvement is more pronounced when using SOWB as the sentiment classifier; the sentiment fairness metric improved by 3.68% when using diversified data for fine-tuning compared to using only the original posts. This trend is also evident in the



(a) Gemma + BERT (b) LLaMA + BERT

(c) Gemma + SOWB (d) LLaMA + SOWB

Figure 5. Comparison of sentiment distributions using ADSD visualization. The *x*-axis represents sentiment scores derived from sentiment analysis models (BERT in subfigures (a) and (b), and SOWB in subfigures (c) and (d)). The *y*-axis, labelled "Density," indicates the probability density of these sentiment scores. Each density plot is normalized such that the total area under the curve equals 1, illustrating how sentiment scores are distributed across the score range. The Gemma and LLaMA models shown in the plots were fine-tuned using the counterfactual method (as illustrated in Figure 1). The ADSD quantifies the degree of sentiment disparity by integrating the absolute difference between the two density curves. A lower ADSD value indicates greater alignment between distributions, suggesting higher fairness in the model's responses.

Table 7. The comparison of ADSD.

Model	Fine-Tuned Data	Classifier	ADSD	Improvement
GPT-2	Original	BERT	0.5527	1.48%
	Counterfactual	BERT	0.5379	
	Original	SOWB	0.2410	1.38%
	Counterfactual	SOWB	0.2272	
Gemma	Original	BERT	0.7304	1.46%
	Counterfactual	BERT	0.7158	
	Original	SOWB	0.3102	3.68%
	Counterfactual	SOWB	0.2734	
LLaMA	Original	BERT	0.8261	2.22%
	Counterfactual	BERT	0.8039	
	Original	SOWB	0.1983	2.65%
	Counterfactual	SOWB	0.1718	

Note: "Original" refers to models fine-tuned using only the initial dataset. "Counterfactual" refers to models fine-tuned using both the initial and counterfactual posts. "BERT" and "SOWB" are the two sentiment classifiers used in Section 3.4.1. "ADSD" is the proposed fairness metric described in Section 3.4.2, where lower values indicate better fairness. The "Improvement" column shows the reduction in ADSD after counterfactual fine-tuning, where a positive difference reflects decreased bias and thus increased fairness.

LLaMA model, where both sentiment classifiers demonstrate that incorporating counterfactual data for LLM fine-tuning can help LLMs achieve fairer sentiment responses.

4.4 Explainable Analysis

To provide an explainable analysis of the LLMs' generated responses, we conducted an explainable analysis of the generated responses of three models using TIGERSCORE. TIGERSCORE provides a detailed analysis of the generated posts, including error location, severity, and penalty scores. Table 8 provides an example of the error analysis of generated responses from Gemma and LLaMA, with the original posts taken from Table 4.

From the table, it is clear that the results from both models might be lacking in terms of accuracy, comprehension, and informativeness. The detailed explainable analysis results show specific areas where each model's generated text is insufficient. For instance, in example 1, the text produced by Gemma faces issues related to comprehensiveness; the explanation shows that this response by Gemma is focused on providing a solution instead of explaining why students might struggle with problem-solving (shown in Table 4, example 1). TIGERSCORE give this a major score reduction of 4 points. For example 2, the responses generated by Gemma and LLaMA show three different issues: accuracy, informativeness, and comprehension. In Gemma's generated response, the accuracy issue is that the response does not address the student's concern about the art teacher's negative attitude; instead, it focuses on proving the teacher wrong. This is a major issue recognized by TIGERSCORE. The informativeness issue refers to the generated response not explaining the mentioned activity, which is a minor issue and causes a score reduction of 2 points. LLaMA's generated response has a major comprehension issue as it is not related to the context in the original post. For example 3, there is an obvious difference in the performance of Gemma and LLaMA3. While the Gemma-generated response has no error recognized by TIGERSCORE, LLaMA3 has three minor accuracy issues, not to mention the specific ways, the importance of number sense, and the speed at which students can solve problems when using number sense. These three cause a total of 3 points reduced. In comparison, using TIGERSCORE, the results generated by LLaMA performed better, with a total reduced score of 7, and Gemma had a reduced score of 10.

We aggregated all TIGERSCORE results, as shown in Figure 6. The average score reductions for GPT-2, Gemma, and LLaMA were 6.23, 1.09, and 0.97, respectively, with a lower score reduction indicating a better AI-generated response. From Figure 6, we observe that GPT-2's generated responses mostly experienced a score reduction of 4 or 8 points (>80%), while Gemma and LLaMA performed better, with a higher proportion of responses showing no score reduction (nearly 20%) or only a 1-point reduction (>70%). Based on the explainable analysis results, we make the following conclusions: (1) Using explainable metrics to evaluate large-scale AI-generated texts is feasible. (2) Overall, the text quality generated by Gemma and LLaMA3 is superior to that of GPT-2, according to the explainable analysis. (3) Although current LLMs offer significant opportunities for online learning support, the quality of this support (e.g., accuracy and comprehension) still requires improvement.

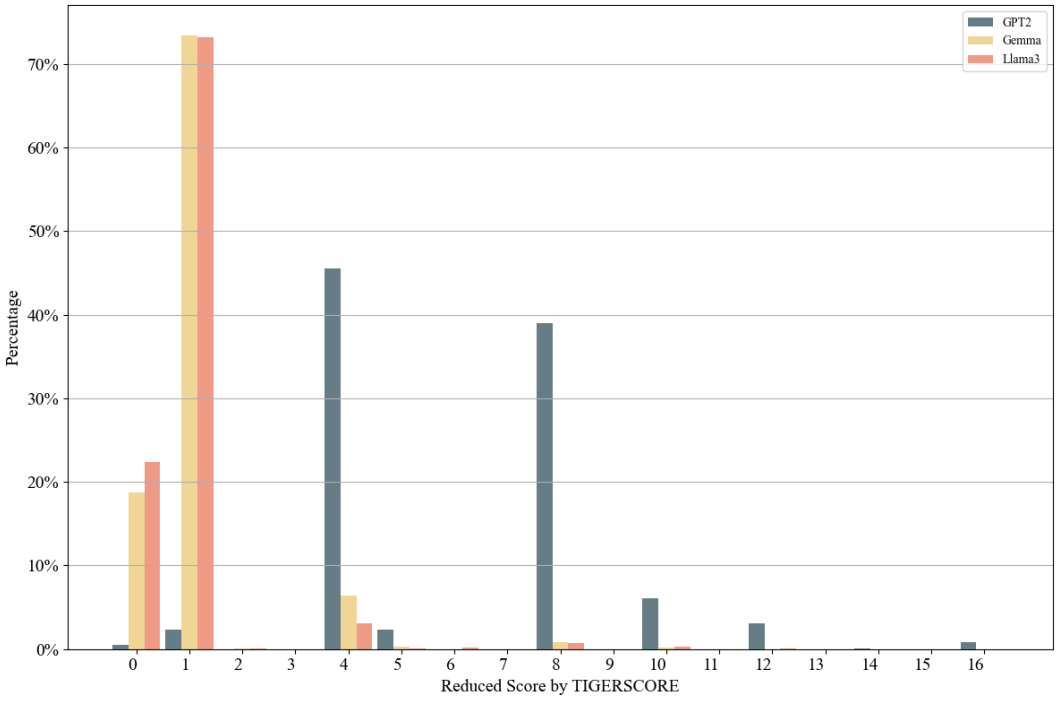


Figure 6. Score reduction distribution by TIGERSCORE.

Table 8. Explainable analysis results.

Original Post	Model	Error Number	Location	Aspect	Explanation	Severity	Score Reduction
1	Gemma	1	Students are often taught to memorize the steps of the process, rather than understanding the process.	Comprehension	The error is a misinterpretation of the given instruction. The instruction was to provide reasons why students might struggle with problem-solving, but the output provides a solution to the problem instead. The correct output should have focused on the reasons for the struggle, not the solution.	Major	4
	LLaMA3	/	/	/	/	/	/
2	Gemma	1	I told them that I was going to prove him wrong.	Accuracy	The output does not address the concern of the student about the language arts teacher's negative attitude towards math. Instead, it focuses on proving the language arts teacher wrong, which may not be the student's main concern. To address the student's concern, the response should focus on how to help the language arts teacher understand that everyone can learn math, and that it is important for teachers to have a positive attitude towards all subjects.	Major	4
		2	I had them do this activity.	Informativeness	The output does not provide any information about the activity that was done in the classroom. To be more informative, the response should describe the activity that was done in the classroom, and how it helped the students understand that everyone can learn math.	Minor	2
	LLaMA3	1	In this paper, we propose a novel method for teaching students.	Comprehension	The output is not related to the given context. It talks about a novel method for teaching students how to prove theorems, which is not relevant to the context. The output should provide advice on how to help the language arts teacher understand the importance of math and how to improve his teaching skills.	Major	4
3	Gemma	/	/	/	/	/	/
	LLaMA	1	Students can use number sense to give answers in many ways.	Accuracy	The output does not mention the specific ways in which students can use number sense to give answers, as mentioned in the source. To improve accuracy, the output should include examples of how students can use number sense, such as recognizing patterns, understanding relationships between numbers, and using mental math strategies.	Minor	1
		2	It is very good for their brain.	Accuracy	The output does not mention the importance of number sense for students, as mentioned in the source. To improve accuracy, the output should include the benefits of number sense for students, such as improving problem-solving skills and enhancing mathematical understanding.	Minor	1
		3	They can solve the problems faster than before.	Accuracy	The output does not mention the speed at which students can solve problems using number sense, as mentioned in the source. To improve accuracy, the output should include the benefits of number sense in terms of problem-solving speed, such as reducing the need for counting or memorization.	Minor	1

Note: "/" means that TIGERSCORE provided feedback indicating that "The model-generated output contains 0 errors, with a total score reduction of 0."

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

5. Discussion and Implications

LLMs, as advanced AI technologies, show significant promise in offering automated support in online learning environments (H. Li et al., 2024). These models, trained on vast amounts of text data, can understand and generate human-like responses. In online learning settings, AI-generated replies from LLMs can provide substantive support, which, in turn, may improve student engagement and academic outcomes (Moore et al., 2019; Du et al., 2023; C. Li & Xing, 2021). However, few studies have systematically examined the sentiment patterns of such responses or the biases that may emerge in emotionally supportive contexts (Huang et al., 2019; Fryer et al., 2022). These biases often originate from the training data itself. LLMs are typically trained on large-scale web-based corpora that reflect real-world language use—language that frequently contains implicit and explicit societal biases related to gender, race, nationality, and other attributes (T. Sun et al., 2019; Sheng et al., 2019; Huang et al., 2019). As a result, LLMs can internalize and reproduce these patterns during generation, sometimes yielding sentiment responses that are uneven or skewed across different demographic groups (C. Li et al., 2022). Rather than alleviating emotional distress, such biased responses risk reinforcing negative experiences for vulnerable learners.

To address this gap, this study aims to explore the sentiment patterns in LLM-generated responses when providing automatic support in online learning environments. While previous research has shown that LLM-generated responses can provide a certain level of emotional support, no studies have yet investigated the specific sentiment patterns present in these responses and whether there is any sentiment bias. We aim to identify and mitigate sentiment bias in LLM-generated responses to ensure that they are both fair and explainable. Using MOOC Posts data, we selected three state-of-the-art models for pre-training and applied counterfactual fine-tuning methods to expand the training dataset, effectively reducing sentiment bias in the models. Our experiments focused on gender attributes, and we introduced a metric called ADSD (shown in Figure 5) to measure sentiment fairness. A counterfactual fine-tuning approach was also proposed to mitigate the sentiment bias in the LLM-generated responses. The results demonstrate that our method effectively reduces sentiment bias in LLM-generated support. Additionally, we conducted an explainable error analysis to evaluate the responses generated by LLMs, offering a scalable approach that can be applied to other contexts. In the discussion, we reflect on our findings; answer our research questions; and assess the performance of different models in educational settings, as well as the broader implications of using LLMs in online learning environments.

5.1 RQ1: What Are the Sentiment Patterns of Responses Generated by LLMs in Online Discussion Forums?

To address RQ1, we conducted a sentiment analysis comparing the responses generated by LLMs with those in the original human-generated dataset. Overall, the LLM-generated content demonstrates the models' ability to either tentatively respond to student posts or articulate relevant information (see examples in Table 4). From a sentiment perspective, as shown in Figure 4, the overall distribution of sentiments—negative, neutral, and positive—in LLM-generated responses (subfigures (b)–(d)) differs notably from that of human responses. Specifically, Gemma and LLaMA tend to produce more neutral responses, whereas GPT-2 exhibits a greater proportion of negative responses. Subsequent t-tests conducted using two sentiment classification models revealed statistically significant differences in sentiment between human- and LLM-generated responses ($p < 0.001$).

Although a previous study by Du and colleagues (2023) using human evaluation reported that LLM-generated responses demonstrated an emotionally supportive reply rate of 58.23%, it found no significant difference compared to human-generated replies. Our study contributes an automated and objective perspective by employing two sentiment analysis models, BERT and SOWB, to evaluate sentiment patterns. This approach provides more fine-grained insights into the emotional tone of LLM-generated content. Another study also relied on human evaluation to assess emotional support, but it included only 150 posts and similarly reported no significant difference between human and LLM-generated replies (C. Li & Xing, 2021). In contrast, we analyzed a much larger dataset of 8,322 posts and applied state-of-the-art LLMs. Our findings indicate that responses generated by Gemma and LLaMA3-8B tend to exhibit a more neutral emotional tone than human replies. To the best of our knowledge, this is the first study to systematically analyze sentiment patterns in LLM-generated responses and directly compare them to the sentiment levels in student-generated replies.

From the sentiment analysis, it is evident that the sentiment patterns of responses generated by LLMs in online discussion forums vary depending on the specific model used. Overall, Gemma and LLaMA tend to produce more neutral responses, whereas GPT-2 exhibits a stronger tendency toward negative sentiment. This increased neutrality suggests that while LLMs can participate in discussions, they often avoid emotionally charged or highly opinionated content—potentially as a safeguard to minimize bias or conflict (C. Li et al., 2022). In contrast, human responses demonstrate a broader range of emotional engagement, often leaning toward more positive or supportive tones, particularly when addressing student queries (as shown in Figure 4 (a)). This divergence between LLM- and human-generated responses highlights a key distinction: while LLMs can replicate factually accurate and contextually appropriate content, they often lack the emotional nuance or empathy commonly exhibited by human participants in online forums. The tendency of models like Gemma and LLaMA to produce neutral responses may stem from training objectives that discourage extreme sentiment polarities (Gemma Team et al., 2024; AI@Meta, 2024). However, this neutrality may also constrain LLMs' ability to provide empathetic support (an essential component of

effective communication) in educational settings (Özhan & Kocadere, 2020; Yang et al., 2022).

5.2 RQ2: Do LLMs Exhibit Sentiment Bias When Generating Responses to Support Online Learning Discussions?

To measure whether LLMs generate sentiment-biased content toward posts of different genders, we used BERT and SOWB to compare sentiment distribution differences between original and counterfactual posts. The results shown in Figure 5 reveal observable differences in sentiment score distribution across various datasets, suggesting the presence of gender bias in the generated responses. This finding aligns with previous research (Zhao et al., 2018; Rudinger et al., 2018; Kiritchenko & Mohammad, 2018). LLMs generate words based on context, and if gender-specific terms frequently appear in the training data, it can lead to gender bias (Caliskan et al., 2017). Additionally, learned from the biased data, LLMs may produce more positive or negative emotional responses in certain contexts associated with gender or content themes, further contributing to sentiment bias (Sheng et al., 2019; Huang et al., 2019).

Inspired by the work of Gardner and colleagues (2019) and Huang and colleagues (2019), we propose a sentiment fairness metric called absolute distributional sentiment divergence (ADSD) to assess fairness across different counterfactual scenarios. ADSD provides a nuanced approach to evaluating sentiment bias in LLM-generated responses. Unlike traditional metrics that focus solely on classification rates (Mehrabi et al., 2021; Gardner et al., 2019), ADSD captures the overall sentiment distribution by comparing the absolute area between two distributions' KDE curves, offering a more holistic view of potential biases. The ADSD values calculated using BERT were notably larger than those calculated by SOWB (as shown in Table 7). For instance, in the LLaMA model, the ADSD value between original and counterfactual posts was 0.8261, while the value measured by SOWB was 0.1983. This discrepancy may stem from BERT's deep learning architecture, which can introduce bias during measurement. Despite these differences, both models indicate bias in LLM-generated responses.

Previous studies have demonstrated that gender differences in cognitive and social presence in online courses significantly affect students' perceived learning and course satisfaction, suggesting that accounting for gender in course design can enhance learning experiences (Cho et al., 2022). Therefore, adopting methods to mitigate gender bias in LLMs is essential to ensuring a fair and inclusive educational environment, as recommended by other studies (C. Li et al., 2022; Kotek et al., 2023; Cho et al., 2022). Addressing bias in AI is crucial for preventing the reinforcement of harmful stereotypes and for ensuring that AI systems deliver fair, unbiased outcomes across diverse populations (Riazy et al., 2020). This study contributes to the analysis of sentiment bias in LLMs within educational settings and draws attention to the application of LLMs in online learning environments. The sentiment metric proposed in this study is extendable to various contexts, including evaluating bias across demographic attributes such as age or race, as well as across multiple attributes simultaneously. It is also applicable to other deep learning models and LLMs, providing a broader framework for assessing model fairness. Additionally, this approach can be adapted to different domains, including healthcare and legal AI systems, where fairness and unbiased decision-making are critical considerations (Hardt et al., 2016; Dressel & Farid, 2018).

5.3 RQ3: How Can We Ensure Fair and Explainable Support for Online Learners in Large-Scale Discussion Forums?

To mitigate sentiment bias in the automatic responses and ensure fairer support for online learners, we conducted experiments using counterfactual posts from ChatGPT 4 API to fine-tune the LLMs. The comparison of sentiment fairness in Table 7 shows that this counterfatual fine-tuning approach helps reduce bias toward gender attributes. By incorporating counterfactual examples, we encourage the model to learn more balanced sentiment distributions, which contribute to fairer interactions with learners of various gender identities. Moreover, reducing bias not only improves learner satisfaction but also promotes an inclusive learning environment. Our approach highlights the potential of counterfactual fine-tuning as an effective strategy for enhancing fairness in AI-driven educational tools. Further research could explore the impact of this method across other demographic attributes and multiple sensitive attributes, such as race and age, to ensure comprehensive fairness in AI-generated content (Caliskan et al., 2017; Liu, Jiao, et al., 2024).

The use of TIGERSCORE in our study directly supports RQ3 by addressing the explainability aspect of LLM-generated responses. The introduction of explainable metrics, such as TIGERSCORE (Jiang et al., 2023), is pivotal in assessing and understanding the utility of large-scale responses generated by LLMs in educational settings. Previous studies evaluated AI-generated text's readability using the FKL (Du et al., 2023) and word perplexity (C. Li & Xing, 2021); however, they only assessed a small set of texts due to the time-consuming nature of manual scoring. The growing emphasis on model interpretability has led to an increase in research on explainable metrics (Zhong et al., 2022; Fu et al., 2024; Liu et al., 2025). This study contributes to this line of research by incorporating explainable analysis into LLM-generated content, demonstrating the potential of such metrics for large-scale educational evaluation. Explainable metrics not only assess readability but also capture the logical coherence of LLM-generated responses. This capability helps refine LLM output and ensures that the generated content is both accurate and pedagogically valuable. Educators can use these insights to scaffold learning more effectively and design interventions that respond to students' needs in online forums.

Despite the potential shown by these technologies, there are significant limitations in the texts generated by current LLMs, 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 required. In this study, we identified various levels of errors made by LLMs, with results indicating that LLaMA generated better outcomes with fewer errors. This may be due to two reasons: firstly, LLaMA3-8B is a more complex model than Gemma, enabling it to learn better and perform more effectively with MOOC Posts data (Touvron et al., 2023); secondly, TIGERSCORE itself is based on the LLaMA series model, which may favour models from the same series during evaluation. Future research should focus on this point and use a diverse set of explainable and automatic evaluation metrics for analysis.

5.4 Implications for Research and Practice

For research, this study introduces the ADSD metric as a novel approach to assessing sentiment fairness in LLM outputs. ADSD can be generalized to evaluate fairness across multiple sensitive attributes beyond gender, such as occupation or geographical location. This provides a flexible and transferable tool for future research on fairness in educational AI applications. Additionally, the integration of fairness and explainability contributes to ongoing discourse on ethical AI in education. While prior studies have highlighted the benefits of LLM-generated support in online learning environments, our findings further emphasize the importance of fair support. We demonstrate that counterfactual training can effectively improve sentiment fairness in LLM responses. This method also has potential applicability in other domains where fairness in automated feedback is critical. Future studies can build on this work by examining how sentiment-aware feedback influences student engagement, persistence, and learning outcomes across diverse learning contexts.

For educators, instructional designers, and AI developers, our findings underscore the importance of emotional tone in AI-generated feedback. In particular, we found that newer LLMs such as Gemma and LLaMA tend to produce more neutral emotional tones. While neutrality may reduce bias, it could also limit emotional resonance with learners. Therefore, curriculum designers and platform developers may consider fine-tuning these models to better align with students' emotional needs and communication preferences. Our study provides actionable insights into the responsible integration of LLMs in online learning environments and highlights the importance of balancing neutrality with empathetic communication.

6. Conclusion, Limitations, and Future Directions

In conclusion, while the advanced capabilities of LLMs such as Gemma and LLaMA3 offer promising opportunities for enhancing interactive learning, their integration into educational frameworks must be undertaken with careful consideration of their limitations, as well as a strong emphasis on ethical implications and educational relevance. This study's analysis and mitigation of sentiment bias in AI-generated content aim to maximize the benefits of these technologies while protecting the learning environment from potential negative effects.

The main contributions of this study are as follows:

1. It examines and compares the sentiment of the replies generated by LLMs with that of human-generated responses. The results indicate that the emotional tone produced by the more recent models, Gemma and LLaMA3, tends to be more neutral than human responses, revealing notable differences in sentiment expression.

2. It introduces a novel sentiment fairness metric, ADSD, which quantifies fairness in sentiment across groups. This metric can be readily applied to various sensitive attributes and contexts. The analysis shows that all three LLMs exhibit a certain degree of gender-related sentiment bias.

3. It applies an explainable and reference-free evaluation metric to assess the quality of AI-generated responses, enabling detailed error analysis. The findings reveal that the responses generated by Gemma and LLaMA3 outperform those of GPT-2 in terms of overall quality, although limitations remain regarding accuracy and comprehension. Overall, this work advances understanding of the practical limitations and potential of LLMs in educational settings and contributes to the development of more equitable and transparent AI-supported learning environments.

Considering the limitations and future objectives identified in this study, several challenges and directions for future research arise. Firstly, the dataset used to fine-tune the model is smaller than typical training sets; however, the methods proposed in this paper can be applied to other scenarios. For example, the sentiment bias analysis can be generalized to other sensitive attributes, such as occupation and geographical location (as shown in Table 3). Moreover, analyzing sentiment bias in human educators and using it as a baseline would facilitate a more meaningful comparison between replies generated by LLMs and those provided by human teachers. Secondly, our reliance on artificially generated counterfactual posts using Chat-GPT 4 API for comparison experiments and model fine-tuning can be resource-intensive, although it is one of the most common approaches used for generating counterfactual texts (e.g., Huang et al., 2019; Fryer et al., 2022). Additionally, our study focuses solely on

gender as a binary sensitive attribute. In future work, we plan to explore open-source tools for generating counterfactuals; test additional sensitive attributes such as race, occupation, and geographic location; and apply our proposed method to multiple sensitive attributes. Furthermore, we employed only one explainable metric, TIGERSCORE, which is based on LLaMA models. This may result in a better score for LLaMA3 than for Gemma. Other evaluation metrics, such as GPTScore (Fu et al., 2024), can also be used to assess generated responses. Since not all AI-generated texts are thoroughly reviewed by humans, errors and biases may remain undetected, potentially affecting output quality. Future research should incorporate more manual analysis focusing on the presence of potential sentiment biases and human satisfaction ratings.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work is supported by the Learning Engineering Virtual Institute under the grant number G-23-2137070 and Jaffe Foundation under the grant number AGR00026932. 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 funding agency.

References

Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. M. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education*, *23*(2), 82–103. https://doi.org/10.1007/s12528-011-9043-x

AI@Meta. (2024). Llama 3 model card [Accessed: 2024-05-30]. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Almatrafi, O., & Johri, A. (2022). Improving MOOCs using information from discussion forums: An opinion summarization and suggestion mining approach. *IEEE Access*, *10*, 15565–15573. https://doi.org/10.1109/ACCESS.2022.3149271

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/http://dx.doi.org/10.2139/ssrn.4337484

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. In J. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 379–396). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Baltà-Salvador, R., Olmedo-Torre, N., Peña, M., & Renta-Davids, A.-I. (2021). Academic and emotional effects of online learning during the COVID-19 pandemic on engineering students. *Education and Information Technologies*, *26*(6), 7407–7434. https://doi.org/10.1007/s10639-021-10593-1

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (ACL 2006), 6 April 2006, Trento, Italy (pp. 313–320). Association for Computational Linguistics. https://doi.org/10.1.1.60.8276

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020), 5–10 July 2020, online (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C. M., Romero-Hall, E., Koutropoulos, A., … Jandrić, P. (2023). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, *18*(1), 53–130. https://doi.org/10.5281/zenodo.7636568

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cercas Curry, A., & Rieser, V. (2018). #MeToo Alexa: How conversational systems respond to sexual harassment. In M. Alfano, D. Hovy, M. Mitchell, & M. Strube (Eds.), *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing* (ACL 2018), 5 June 2018, New Orleans, Louisiana, USA (pp. 7–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0802

Cho, M., Lim, S., Lim, J., & Kim, O. (2022). Does gender matter in online courses? A view through the lens of the community of inquiry. *Australasian Journal of Educational Technology*, *38*(6), 169–184. https://doi.org/10.14742/ajet.7194

Cleveland-Innes, M., & Campbell, P. (2012). Emotional presence, learning, and the online learning environment. *International Review of Research in Open and Distributed Learning*, *13*(4), 269–292. https://doi.org/10.19173/irrodl.v13i4.1234

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Dickson-Deane, C., & Chen, H.-L. (2018). Understanding user experience. In D. Mehdi Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (4th ed., pp. 7599–7608). IGI Global. https://doi.org/10.4018/978-1-5225-2255-3.ch661

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Du, H., & Xing, W. (2023). Leveraging explainability for discussion forum classification: Using confusion detection as an example. *Distance Education*, *44*(1), 190–205. https://doi.org/10.1080/01587919.2022.2150145

Du, H., Xing, W., & Pei, B. (2023). Automatic text generation using deep learning: Providing large-scale support for online learning communities. *Interactive Learning Environments*, *31*(8), 5021–5036. https://doi.org/10.1080/10494820.2021.1993932

Dumford, A. D., & Miller, A. L. (2018). Online learning in higher education: Exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education*, *30*(3), 452–465. https://doi.org/10.1007/s12528-018-9179-z

Froehlich, L., & Weydner-Volkmann, S. (2024). Adaptive interventions reducing social identity threat to increase equity in higher distance education: A use case and ethical considerations on algorithmic fairness. *Journal of Learning Analytics*, *11*(2), 112–122. https://doi.org/10.18608/jla.2024.8301

Fryer, Z., Axelrod, V., Packer, B., Beutel, A., Chen, J., & Webster, K. (2022). Flexible text generation for counterfactual fairness probing. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, & Z. Talat (Eds.), *Proceedings of the Sixth Workshop on Online Abuse and Harms* (WOAH 2022), 14 July 2022, Seattle, Washington, USA (hybrid) (pp. 209–229). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.20

Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2024). GPTScore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*. https://doi.org/10.48550/arXiv.2302.04166

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 225–234). ACM. https://doi.org/10.1145/3303772.3303791

Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., … Kenealy, K. (2024). Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*. https://doi.org/10.48550/arXiv.2403.08295

Gottipati, S., Shankararaman, V., & Ramesh, R. (2019). TopicSummary: A tool for analyzing class discussion forums using topic based summarizations. In *2019 IEEE Frontiers in Education Conference* (FIE 2019), 16–19 October 2019, Cincinnati, Ohio, USA (pp. 1–9). IEEE. https://doi.org/10.1109/FIE43999.2019.9028526

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS 2016), 5–10 December 2016, Barcelona, Spain (pp. 3323–3331). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3157382.3157469

Hernández-Sellés, N., Pablo-César Muñoz-Carril, & González-Sanmamed, M. (2019). Computer-supported collaborative learning: An analysis of the relationship between interaction, emotional support and online collaborative tools. *Computers & Education*, *138*, 1–12. https://doi.org/10.1016/j.compedu.2019.04.012

Hew, K. F., & Cheung, W. S. (2012). *Student participation in online discussions: Challenges, solutions, and future research*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-2370-6

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724. https://doi.org/10.1016/j.compedu.2019.103724

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2004), 22–25 August 2004, Seattle, Washington, USA (pp. 168–177). ACM. https://doi.org/10.1145/1014052.1014073

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., & Kohli, P. (2019). Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*. https://doi.org/10.48550/arXiv.1911.03064

Hwang, G.-J., & Chen, N.-S. (2023). Exploring the potential of generative artificial intelligence in education: Applications, challenges, and future research directions. *Educational Technology & Society*, *26*(2). https://drive.google.com/file/d/15zj51LzLpsE-LE-04WMycl4uvFTOn8x-/view

Jiang, D., Li, Y., Zhang, G., Huang, W., Lin, B. Y., & Chen, W. (2023). TIGERScore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*. https://doi.org/10.48550/arXiv.2310.00752

Khalil, M., Prinsloo, P., & Slade, S. (2023). Fairness, trust, transparency, equity, and responsibility in learning analytics. *Journal of Learning Analytics*, *10*(1), 1–7. https://doi.org/10.18608/jla.2023.7983

Kim, D., Park, Y., Yoon, M., & Jo, I. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education*, *30*, 30–43. https://doi.org/10.1016/J.IHEDUC.2016.03.002

Kiritchenko, S., & Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In M. Nissim, J. Berant, & A. Lenci (Eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (SEM 2018), 5–6 June 2018, New Orleans, Louisiana, USA (pp. 43–53). Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2005

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174–202). Routledge. https://doi.org/10.4324/9780429329067

Kotek, H., Dockum, R., & Sun, D. Q. (2023). Gender bias and stereotypes in large language models. In M. Bernstein, S. Savage, & A. Bozzon (Eds.), *Proceedings of The ACM Collective Intelligence Conference* (CI 2023), 6–9 November 2023, Delft, Netherlands (pp. 12–24). ACM. https://doi.org/10.1145/3582269.3615599

Langford, C. P. H., Bowsher, J., Maloney, J. P., & Lillis, P. P. (1997). Social support: A conceptual analysis. *Journal of Advanced Nursing*, *25*(1), 95–100. https://doi.org/10.1046/j.1365-2648.1997.1997025095.x

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education*, *31*, 186–214. https://doi.org/10.1007/s40593-020-00235-x

Li, C., Xing, W., & Leite, W. (2022). Building socially responsible conversational agents using big data to support online learning: A case with Algebra Nation. *British Journal of Educational Technology*, *53*(4), 776–803. https://doi.org/10.1111/bjet.13227

Li, H., Li, C., Xing, W., Baral, S., & Heffernan, N. (2024). Automated feedback for student math responses based on multi-modality and fine-tuning. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 763–770). ACM. https://doi.org/10.1145/3636555.3636860

Li, L., Johnson, J., Aarhus, W., & Shah, D. (2022). Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. *Computers & Education*, *176*, 104354. https://doi.org/10.1016/j.compedu.2021.104354

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. *arXiv preprint arXiv:2106.13219*. https://doi.org/10.48550/arXiv.2106.13219

Liu, Z., Jiao, X., Li, C., & Xing, W. (2024). Fair prediction of students' summative performance changes using online learning behavior data. In D. Joyner, B. Paaßen, & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (EDM 2024), 14–17 July 2024, Atlanta, Georgia, USA (pp. 686–691). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729918

Liu, Z., Xing, W., Jiao, X., & Li, C. (2025). What are the differences between student and ChatGPT-generated pseudocode? Detecting AI-generated pseudocode in high school programming using explainable machine learning. *Education and Information Technologies*, *30*, 14853–14892. https://doi.org/10.1007/s10639-025-13385-z

Liu, Z., Xing, W., & Li, C. (2024). Explainable analysis of AI-generated responses in online learning discussions. In D. Joyner, B. Paaßen, & C. D. Epp (Eds.), *Proceedings of the Educational Data Mining 2024 Workshop: Leveraging Large Language Models for Next-Generation Educational Technologies* (EDM 2024), 14–17 July 2024, Atlanta, Georgia, USA. Educational Data Mining Society. https://doi.org/10.13140/RG.2.2.24309.38881/1

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In V. Nigam, T. B. Kirigin, C. Talcott, J. Guttman, S. Kuznetsov, B. T. Loo, & M. Okada (Eds.), *Logic, language, and security. Lecture notes in computer science* (pp. 189–202, Vol. 12300). Springer. https://doi.org/10.1007/978-3-030-62077-6_14

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6). https://doi.org/10.1145/3457607

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 100152. https://doi.org/10.1016/j.caeai.2023.100152

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: Examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, *27*(5–6), 655–669. https://doi.org/10.1080/10494820.2019.1610453

Naseer, F., Khalid, M. U., Ayub, N., Rasool, A., Abbas, T., & Afzal, M. W. (2024). Automated assessment and feedback in higher education using generative AI. In R. C. Sharma & A. Bozkurt (Eds.), *Transforming education with generative AI: Prompt engineering and synthetic content creation* (pp. 433–461). IGI Global. https://doi.org/10.4018/979-8-3693-1351-0.ch021

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*. https://doi.org/10.48550/arXiv.1103.2903

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (WWW 2016), 11–15 April 2016, Montréal, Québec, Canada (pp. 145–153). ACM. https://doi.org/10.1145/2872427.2883062

Onan, A. (2021). Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*(3), 572–589. https://doi.org/10.1002/cae.22253

Özhan, Ş. Ç., & Kocadere, S. A. (2020). The effects of flow, emotional engagement, and motivation on success in a gamified online learning environment. *Journal of Educational Computing Research*, *57*(8), 2006–2031. https://doi.org/10.1177/0735633118823159

Parmar, D., Dewan, M. A. A., & Wen, D. (2023). Automatic analysis of online course discussion forum: A short review. In *2023 IEEE Canadian Conference on Electrical and Computer Engineering* (CCECE 2023), 24–27 September 2023, Regina, Saskatchewan, Canada (pp. 210–215). IEEE. https://doi.org/10.1109/CCECE58730.2023.10289065

Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024). The ultimate guide to fine-tuning LLMs from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*. https://doi.org/10.48550/arXiv.2408.13296

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In H. C. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *Proceedings of the 12th International Conference on Computer Supported Education* (CSEDU 2020), 2–4 May 2020, online (pp. 15–25, Vol. 1). SciTePress. https://doi.org/10.5220/0009324100150025

Roundtree, A. K. (2023). AI explainability, interpretability, fairness, and privacy: An integrative review of reviews. In H. Degen & S. Ntoa (Eds.), *Artificial intelligence in HCI. HCII 2023. Lecture notes in computer science* (pp. 262–282, Vol. 14050). Springer. https://doi.org/10.1007/978-3-031-35891-3_19

Rovai, A. P. (2007). Facilitating online discussions effectively. *The Internet and Higher Education*, *10*(1), 77–88. https://doi.org/10.1016/J.IHEDUC.2006.10.001

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers)* (NAACL HLT 2018), 1–6 June 2018, New Orleans, Louisiana, USA (pp. 8–14, Vol. 2). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2002

Sabbaghi, S. O., Wolfe, R., & Caliskan, A. (2023). Evaluating biased attitude associations of language models in an intersectional context. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, & A. John (Eds.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES 2023), 8–10 August 2023, Montréal, Québec, Canada (pp. 542–553). ACM. https://doi.org/10.1145/3600211.3604666

Sahay, A., Gholkar, S., & Arya, K. (2019). Selection-based question answering of an MOOC. *arXiv preprint arXiv:1911.07629*, 1–5. https://doi.org/10.48550/arXiv.1911.07629

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL 2019), 28 July–2 August 2019, Florence, Italy (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020), 5–10 July 2020, online (pp. 5248–5264). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)* (ACL-IJCNLP 2021), 1–6 August 2021, online (pp. 4275–4293, Vol. 1). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.330

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), 3–7 November 2019, Hong Kong, China (pp. 3407–3412). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1339

Shumaker, S. A., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of Social Issues*, *40*(4), 11–36. https://doi.org/10.1111/j.1540-4560.1984.tb01105.x

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, *17*, 1920–1930. https://doi.org/10.1109/TLT.2024.3396873

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL 2019), 28 July–2 August 2019, Florence, Italy (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Sun, Y., & Gao, F. (2017). Comparing the use of a social annotation tool and a threaded discussion forum to support online discussions. *The Internet and Higher Education.*, *32*, 72–79. https://doi.org/10.1016/J.IHEDUC.2016.10.001

Tang, H., Xing, W., & Pei, B. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, *39*(3), 353–372. https://doi.org/10.1080/01587919.2018.1476841

Tegos, S., Demetriadis, S., & Karakostas, A. (2015). Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education*, *87*, 309–325. https://doi.org/10.1016/j.compedu.2015.07.014

Thoits, P. A. (1986). Social support as coping assistance. *Journal of Consulting and Clinical Psychology*, *54*(4), 416–423. https://doi.org/10.1037/0022-006X.54.4.416

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971

Van De Poel, I. (2021). Design for value change. *Ethics and Information Technology*, *23*(1), 27–31. https://doi.org/10.1007/s10676-018-9461-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS 2017), 4–9 December 2017, Long Beach, California, USA (pp. 6000–6010). https://dl.acm.org/doi/10.5555/3295222.3295349

Vayre, E., & Vonthron, A.-M. (2017). Psychological engagement of students in distance and online learning: Effects of self-efficacy and psychosocial processes. *Journal of Educational Computing Research*, *55*(2), 197–218. https://doi.org/10.1177/0735633116656849

Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H., & Wilson, S. (2023). Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*. https://doi.org/10.48550/arXiv.2302.02463

Wang, B., Shen, T., Long, G., Zhou, T., & Chang, Y. (2021). Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. *arXiv preprint arXiv:2109.02403*. https://doi.org/10.48550/arXiv.2109.02403

Wang, Q., Rose, C. P., Ma, N., Jiang, S., Bao, H., & Li, Y. (2022). Design and application of automatic feedback scaffolding in forums to promote learning. *IEEE Transactions on Learning Technologies.*, *15*(2), 150–166. https://doi.org/10.1109/TLT.2022.3156914

Williamson, J. M. L., & Martin, A. G. (2010). Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. *International Journal of Clinical Practice*, *64*(13), 1824–1831. https://doi.org/10.1111/j.1742-1241.2010.02408.x

Wong, G. K., Li, Y. K., & Lai, X. (2021). Visualizing the learning patterns of topic-based social interaction in online discussion forums: An exploratory study. *Educational Technology Research and Development*, *69*(5), 2813–2843. https://doi.org/10.1007/s11423-021-10040-5

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, *57*(3), 547–570. https://doi.org/10.1177/0735633118757015

Xu, W., Wang, D., Pan, L., Song, Z., Freitag, M., Wang, W. Y., & Li, L. (2023). INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. *arXiv preprint arXiv:2305.14282*. https://doi.org/10.48550/arXiv.2305.14282

Yan, L., Sha, L., Zhao, L., Li, Y., Martínez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yang, G., Sun, W., & Jiang, R. (2022). Interrelationship amongst university student perceived learning burnout, academic self-efficacy, and teacher emotional support in China's English online learning context. *Frontiers in Psychology*, *13*, 829193. https://doi.org/10.3389/fpsyg.2022.829193

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL HLT 2018, Short Papers), 1–6 June 2018, New Orleans, Louisiana, USA (pp. 15–20, Vol. 2). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2003

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., & Luo, Z. (2024). LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Y. Cao, Y. Feng, & D. Xiong (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (System Demonstrations)* (ACL 2024), 11–16 August 2024, Bangkok, Thailand (pp. 400–410, Vol. 3). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-demos.38

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., & Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (AMNLP 2022), 7–11 December 2022, Abu Dhabi, United Arab Emirates (pp. 2023–2038). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.131

Zhou, J., Chen, F., & Holzinger, A. (2022). Towards explainability for AI fairness. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI—Beyond explainable AI* (Vol. 13200). Springer. https://doi.org/10.1007/978-3-031-04083-2_18

Zylich, B., Viola, A., Toggerson, B., Al-Hariri, L., & Lan, A. (2020). Exploring automated question answering methods for teaching assistance. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education. AIED 2020. Lecture notes in computer science* (pp. 610–622, Vol. 12163). Springer. https://doi.org/10.1007/978-3-030-52237-7_49