# Enhancing Trust in Generative AI: Investigating Explainability of LLMs to Analyse Confusion in MOOC Discussions

Yuanyuan Hu[1], Nasser Giacaman[1] and Claire Donald[1]

[1] *The University of Auckland, 5 Grafton Rd, Auckland, 1010, New Zealand*

### Abstract

Providing feedback to address learners' confusion in a personalised and timely manner can enhance learning engagement and deeper understanding in large-scale online courses, particularly Massive Open Online Courses (MOOCs). This goal aligns with a key objective within the Learning Analytics (LA) community. The advent of Generative Artificial Intelligence (GenAI) tools presents the potential to identify learners' confusion in vast numbers of discussion texts and provide automatically-generated and adaptive feedback to learners rapidly. However, a lack of trust in AI-generated content among educators and learners is an obstacle to building effective GenAI-based LA solutions. This paper discusses the potential of enhancing trust in GenAI tools by improving the transparency and explainability of the large language models (LLMs) — a foundation of GenAI. We illustrate this approach through a pilot study where we apply an explainable AI (XAI) method — the Integrated Gradients — to decipher LLM-based predictions regarding learners' confusion in MOOC discussions. The findings suggest promising reliability in the XAI method's ability to identify word-level indicators of confusion in MOOC messages. The paper concludes by advocating the integration of XAI methods in GenAI applications, aiming to foster wider acceptance and efficacy of future GenAI-based LA solutions.

### Keywords

Generative AI, Trust, Explainable AI, Learning Analytics, AIED, XAI, XAI-ED

## 1. INTRODUCTION

Confusion, a common emotion during learning, is often an obstacle for learners to move forward [1]. While a certain level of confusion can encourage learning engagement [2], this confusion may also evolve into frustration and finally lead to boredom without timely interventions [3]. In distance learning contexts, particularly Massive Open Online Courses (MOOCs), low participation and drop-out rates may increase due to the impact of learners' emotions, such as confusion [4].

MOOCs offer high-quality, open-access, rich, online learning resources, and micro-credentials regardless of university-entry barriers, empowering a diverse range of learners to study at their own pace. As learning in MOOCs is entirely virtual and asynchronous, discussion forums become key venues for interaction and communication between learners and instructors. In MOOCs, resolving numerous queries and confusion raised by a huge number of learners in discussion forums is a significant challenge due to the limited availability of educators [5, 6, 7]. Behavioural and physiological measures, such as facial expressions and skin conductance, have successfully discerned learners' confusion in traditional small to medium classrooms [8]. However, these measures are impractical to be implemented in MOOCs. Researchers explore solutions to provide adaptive, immediate responses to address learners' confusion and improve learning engagement in MOOC discussion forums [9]. This objective is also a crucial goal of the Learning Analytics (LA) community [10].

The increasing availability of generative artificial intelligence (GenAI) tools, such as ChatGPT [11] and Gemini [12], has opened fresh possibilities for LA research. Investigating the feasibility of applying GenAI tools in higher education practices has displayed promising outcomes, such as automatic generation of academic writing [13], adaptive responses to discussion-forum posts [9],

Proceedings Acronym: Proceedings Name, Month XX–XX, YYYY, City, Country

y.hu@auckland.ac.nz (Y. Hu); n.giacaman@auckland.ac.nz (N. Giacaman); c.donald@auckland.ac.nz (C. Donald)
0000-0003-2592-8473 (Y. Hu); 0000-0001-6885-1571 (N. Giacaman); 0000-0003-2803-7537 (C. Donald)

automated code review [14], and personalised summary and feedback on students' writing assignments [15]. Despite the opportunities in education, the breakthroughs of GenAI techniques have also sparked debates on their ethical concerns, such as biases and reliability concerns about the texts generated, namely *trust* in GenAI [16, 17, 18]. Also, the EU Parliament calls for safety, transparency, accountability, equality, and eco-friendliness in AI techniques to avoid harmful effects [19].

Deficiency of trust in AI-generated content among educators and learners is an obstacle to developing reliable and effective GenAI-based LA (Gen-LA) solutions for teaching and learning [16, 20, 21]. This issue of deficiency stems from the fundamental architecture of GenAI, specifically the transformer-based large language models (LLMs). A notable challenge faced by LLMs and inherited by GenAI is deep learning models' inability to explain the mechanisms and reasoning in their decision-making processes [22]. Explainable artificial intelligence (XAI) methods can contribute to interpreting obscured 'black-box' mechanisms hiding behind deep learning models [22, 23]. Applying XAI techniques to provide clear rationales in AI-generated content is required for designing and developing trustworthy educational AI systems [24].

Based on these studies, our research interest focuses on investigating the potential of employing XAI methods to decode word-level indicators in the prediction made by LLMs, particularly in identifying learners' confusion in MOOC discussions. Detecting learners' confusion timely and accurately is a prerequisite for providing them with adaptive feedback in MOOC discussions. Earlier studies decoded linguistic cues as indicators of confusion in MOOC discussions using different machine learning and deep learning models [1, 25, 26, 27]. As a preliminary step of an ongoing project, this small project attempts to provide proof of concept for enhancing trust in future GenAI applications by improving the transparency and explainability of LLMs. Thus, the research question in this paper is "What can we gain from using XAI methods to interpret LLMs' predictions of learners' confusion in MOOC discussions?" We assume that XAI methods can discern positive and negative indicators behind LLMs' processes for identifying confusion in MOOC discussions. We conduct a test case in this paper to examine this assumption.

In the following section, we will review the related work that shaped our study: learners' confusion identification in MOOC discussions and applications of XAI methods to interpret important features of confusion predictions. Subsequently, we will illustrate a pilot study to address our research question. Finally, suggestions for improving trust in future GenAI-LA solutions based on the implications of the pilot study will be expounded at the end of this paper.

## 2. RELATED WORK

A pioneering study on the detection of learners' confusion in MOOC discussions is Agrawal et al.'s research [28] that developed a system using bag-of-words, the conversational position of discussion messages, the number of likes, and learners' grades to identify confusion and recommended minute-resolution video clips to learners accordingly. This study also developed the Stanford MOOC discussion data sets, which are used in our pilot study explained in Section 3.1. These data sets were also used in most of the previous studies on analysing learners' confusion in MOOC discussions [1, 26, 27, 29, 30, 31].

After Agrawal et al.'s work [28], Bakharia [30] applied a Support-vector-machine classifier to detect confusion, urgency, and sentiments in MOOC discussions, achieving over 70% accuracy rates in domain-specific courses. Zeng et al. [32] trained an Elastic-net model using content-related features (e.g., readability index, the number of words in a post, topicality, etc.) and community-related features (e.g., the number of likes and reads, etc.) to identify confusion and urgency in MOOC discussions, reaching an over 80% accuracy in specific domain data sets.

Building on the previous work, Atapattu et al. [1, 26] applied a random forest classifier with solely linguistic features to identify learners' confusion in MOOC discussions, improving the accuracy to over 83% $F_1$ score in all domain-specific data sets and to $F_1$ scores between 70.7% to 84.5% in cross-domain data sets. These approaches not only underscore the significant role of linguistic features in identifying learners' confusion within MOOC discussions, but also imply the nuanced language cues that can distinguish confusion messages from non-confusion ones.

Other trials that explore machine learning and deep learning methods to detect confusion in MOOC discussions after Atapattu et al.'s [1, 26] work focused merely on enhancing classification performance rather than deciphering indicators of learners' confusion, such as applying a Transformers classification model in Chanaa and El Faddouli [31] and comparing different machine learning methods in Bhumireddy and Anala [33].

Alrajhi et al. [25] offer a preliminary example of employing XAI methods to interpret the prediction reasons of a Transformers model with ontology methods to detect urgent MOOC discussion messages. While this study offers valuable insights for investigating XAI methods, providing more details about the prediction and interpretation processes would be more helpful for further studies in this area. Du and Xing [27] developed an explainable text classifier framework to identify confusion in MOOC discussions based on a legal services model. However, they mentioned that their work might have limitations in interpreting negative indicators for different levels of confusion.

# 3. A PILOT STUDY

In this section, we demonstrate a pilot study to investigate the research question proposed in Section 1. We will explain the data sets used in this study, the architecture of LLMs-based classifiers for confusion detection, an XAI method employed to interpret word-level indicators for model predictions, and experimental results.

## 3.1. Datasets Description

Data sets used in this study came from discussion posts and replies in the Stanford University public MOOCs, which contain archived runs of eleven courses [28]. These courses involve multiple topics mainly from three disciplines: Education, Medicine, and Humanities. Each message — a discussion post or a reply — was classified by three expert coders independently into degrees of confusion from 1 (extremely knowledgeable) to 7 (extremely confused). In our pilot study, the degree of 4 was regarded as neutral, 1 to 3 as non-confusion, and 5 to 7 as confusion, which was the same categorisation way in Atapattu et al.'s work [1].

Following Atapattu et al.'s [1] work, we trained and tested our binary classification model, which will be explained in the next section, through two experiments. One included the neutral messages in the confusion ones as a 'broader' confusion class. The other excluded these neutral messages in training and testing processes. Both experiments' outcomes will be illustrated in Section 3.3. We pre-processed the text data by expanding abbreviations, eliminating repeated characters and extraneous spaces, and removing messages with less than three words. Table 1 demonstrates the distribution of messages classified in each category across three domain-specific data sets after the data cleaning.

**Table 1**
**Number of messages classified as Non-Confusion, Confusion, and Neutral in Education, Medicine, Humanities, and all three data sets.**

| Set | Non-Confusion | Confusion | Neutral |
|---|---|---|---|
| Education | 6650 | 638 | 2446 |
| Medicine | 1577 | 1587 | 6339 |
| Humanities | 1533 | 2252 | 5872 |
| Total | 9760 | 4477 | 14657 |

## 3.2. Classifier Architecture and the XAI method

We applied and fine-tuned a DistilBERT model — a faster and lightweight transformer-based deep learning model [34] — to predict confusion or non-confusion of messages in MOOC discussions. This pre-trained model has been applied to a broader range of natural language processing solutions, particularly where the implementation of LLMs is not feasible due to hardware resource limitations. The DistilBERT model has achieved excellent performance in sentiment analysis tasks [35]. As a faster and smaller LLM but maintaining a competitive level of accuracy, the DistilBERT model will be a better option for our pilot study on the model explanation process of the automatic confusion analysis to provide concept of concept rather than very large and computationally expensive LLMs, such as GPT-4 [36].

Based on Alrajhi et al.'s work [25], we employed the Integrated Gradients method from the Captum library for PyTorch [37] to gain a deeper understanding of the decision-making processes (i.e., positive, or negative indicators) within our DistilBERT-based confusion classifier. The Integrated Gradients method computes the prediction feature importance by integrating gradients of the deep learning model's outputs (e.g., classes) regarding the inputs (e.g., words and sentences), from non-informative baseline inputs to actual inputs, evaluating each feature's contribution to the prediction output.

## 3.3. Results

### 3.3.1. Classification performance

In training and testing processes of domain-specific sets, our fine-tuned DistilBERT model achieved the best-performing weighted-averaged $F_1$ scores of 0.74, 0.90, 0.87 and 0.83 in the Education, Medicine, Humanities, and all three data sets, respectively, where we regarded neutral messages as confusion messages based on Atapattu et al.'s work [1]. When we excluded these neutral messages in the training step, weighted-averaged $F_1$ scores increased to 0.95, 0.90, 0.90, and 0.92, as summarised in Table 2. Our models reached an average higher performance than random forest classifiers applied in a previous study with and without neutral messages [1]. These results suggest that the neutral messages, which were classified between confusion and non-confusion messages by expert coders, affect the model performance, particularly in the Education set where neutral messages contributed the major percentage.

**Table 2**
**Fine-tuned model classification performance in Education, Medicine, Humanities, and all three data sets.**

| Set | Weighted average $F_1$ (including neutral data) | Weighted average $F_1$ (excluding neutral data) |
|---|---|---|
| Education | 0.74 | 0.94 |
| Medicine | 0.90 | 0.90 |
| Humanities | 0.87 | 0.90 |

### 3.3.2. Word-level indicators for confusion identification

This section presents results from experiments designed to predict confusion and non-confusion in the MOOC messages, where neutral messages were excluded. The reliability of these experiments is underscored by the model's high performance, achieving over 0.90 $F_1$ scores. Due to the page limits of this workshop paper, interpretation samples from the best-performing Education set are displayed in Figure 1 and Figure 2 as examples. We highlight negative indicators in red and positive ones in green. The intensity of the green correlates with the strength of the positive attribution.

While the paper only showcases examples from the Education dataset, we provide a summary of the findings from experiments conducted on domain-specific and all three datasets as follows. Strong word-level indicators to predict MOOC learners' confused messages positively are 1) first-person singular and plural, 2) question stems, 3) question bigrams, 4) confusion expressions, and 5) the question mark. Strong indicators that can predict non-confused messages positively are 1) second-person pronouns and 2) academic writing expressions. These interpretation outcomes by the XAI method strongly align with the indicators found in previous studies [1, 26].

# 4. IMPLICATIONS AND FUTURE WORK

## 4.1. Implications

We can answer our research questions as follows. Outcomes of our pilot study demonstrate promising reliability of using the Integrated Gradients method with the fine-tuned DistilBERT model to discern word-level predictors in the MOOC discussions. This is because indicators of confusion detected in our study are in line with the linguistic indicators identified by the previous studies using tree-based machine learning classifiers [1, 26]. Unlike hidden computation of deep learning models, tree-based machine learning algorithms are often regarded as "white-box" models due to their clear, transparent decision-making rules and easy, straightforward tracking paths of every-step impacts of input features on outputs. This is also the main reason that white-box algorithms can be preferences for educational studies [38]. Robustness of a certain degree can be implied if indicators from the XAI method are similar to the important features from white-box algorithms. Future research can employ XAI methods in tandem with LLMs to enhance the transparency and trustworthiness of deep learning mechanisms, leveraging GenAI-LA solutions to be more accessible and understandable to non-technical audiences.

A possible application of XAI techniques in GenAI-LA solutions is offering clear and UX-friendly designed rationales, along with automatically generated and personalised feedback to urgent MOOC posts. A previous study suggests that GPT2-generated replies to MOOC posts can reach a similar degree of emotional and community support as human tutors although a lower extent of informality [39]. This promising result encourages further studies to investigate the potential of applying GenAI techniques to provide learners with automatic responses in large-scale online learning scenarios. We recommend employing XAI methods to highlight words or phrases that attribute high importance to GenAI's decision-making processes for each part of the responses generated. In this way, learners and educators can gain insights into how AI tools approach their queries, improving their trust in AI-generated content. Also, learners can refine question-posing strategies in discussion forums to elicit accurate responses from GenAI agents according to rationales provided by XAI methods.

XAI methods can also be integrated into other GenAI-LA applications such as AI-assisted writing assessment. A writing analytics tool, AcaWriter, applies XAI designs to offer sentence-level and document-level feedback in learners' academic writing assessments [40]. A recent study indicates that ChatGPT can generate high-quality feedback on summarising topics of students' assignments and providing process-focus suggestions [15]. We assume XAI methods also have the potential to offer distributed rationales at word, sentence, concept, and organisation levels with grading rubrics during GenAI-assisted writing processes. In this way, scores and advice provided by GenAI tools would become more transparent and credible to both learners and educators.

The LA community calls for redefining our perception of learners in the AI era [41, 42]. Learners can gain personalised feedback from GenAI as a new way of learning. At the same time, learners can iteratively coach a GenAI tool to align its responses with their expectations. GenAI is regarded as a full participant in conversational education systems now [43]. With the improvement of transparency and explainability by providing learners with rationales in AI's decision-making mechanisms, they will coach GenAI more easily and effectively for personalised learning demands. This reciprocal learning model, akin to the 'Ako' concept from Māori culture where roles between educator and learner are interchangeable, may offer innovative ways to enhance skills such as problem-solving, collaboration, and self-regulated learning in the AI era.

## 4.2. Limitations and Future Work

This study has two main limitations. Firstly, the pilot study only provides a trial of using an XAI method to explain the positive and negative indicators in confusion predictions of the LLMs-based classifier, which is a vital foundation of GenAI. This XAI method may not be directly extended to GenAI models. Secondly, the LLMs-based classifier for identifying confusion messages was trained and fine-tuned by using discussion data from three domains (i.e., education, medicine, and humanities), which still needs further refinement on MOOC discussions from other domains to improve the model's generalisability.

Our future work will investigate the feasibility of using XAI methods to detect key indicators within learners' queries that result in content generated by GenAI. We will also explore methods to visualise these indicators at a word level in a way that is intuitive and readable for learners and educators. This future research will enhance the feasibility and user-friendliness of GenAI-LA solutions towards human-AI collaboration on teaching and learning processes in the age of AI.

## References

[1] T. Atapattu, K. Falkner, M. Thilakaratne, L. Sivaneasharajah, and R. Jayashanka, "An Identification of Learners' Confusion through Language and Discourse Analysis," *arXiv preprint arXiv:1903.03286*, 2019.

[2] S. D'Mello and A. Graesser, "Confusion and its dynamics during device comprehension with breakdown scenarios," *Acta Psychol (Amst)*, vol. 151, pp. 106–116, 2014, doi: 10.1016/j.actpsy.2014.06.005.

[3] M. K. Chandrasekaran, M.-Y. Kan, B. C. Y. Tan, and K. Ragupathi, "Learning Instructor Intervention from MOOC Forums: Early Results and Issues," Apr. 2015, [Online]. Available: http://arxiv.org/abs/1504.07206

[4] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A Review and Future Research Directions," *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, Jul. 2019, doi: 10.1109/TLT.2018.2856808.

[5] I. Buchem *et al.*, "Integrating Mini-Moocs into Study Programs in Higher Education During Covid-19. Five Pilot Case Studies in Context of the Open Virtual Mobility Project," *Human and Artificial Intelligence for the Society of the Future*, pp. 299–310, 2020, doi: 10.38069/edenconf-2020-ac0028.

[6] H. Cha and H. J. So, *Integration of Formal, Non-formal and Informal Learning Through MOOCs*. Springer Singapore, 2020. doi: 10.1007/978-981-15-4276-3_9.

[7] Y. Hu, C. Donald, and N. Giacaman, "Cross Validating a Rubric for Automatic Classification of Cognitive Presence in MOOC Discussions," *International Review of Research in Open and Distributed Learning*, vol. 23, no. 2, pp. 242–260, 2021, doi: https://doi.org/10.19173/irrodl.v23i3.5994.

[8] A. Arguel, L. Lockyer, O. V. Lipp, J. M. Lodge, and G. Kennedy, "Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments," *Journal of Educational Computing Research*, vol. 55, no. 4, pp. 526–551, 2017, doi: 10.1177/0735633116674732.

[9] C. Li and W. Xing, "Natural Language Generation Using Deep Learning to Support MOOC Learners," *Int J Artif Intell Educ*, vol. 31, no. 2, pp. 186–214, Jun. 2021, doi: 10.1007/s40593-020-00235-x.

[10] "What is Learning Analytics?" Accessed: Jan. 23, 2024. [Online]. Available: https://www.solaresearch.org/about/what-is-learning-analytics/

[11] "ChatGPT." Accessed: Jan. 23, 2024. [Online]. Available: https://chat.openai.com

[12] "Gemini." Accessed: Jan. 23, 2024. [Online]. Available: https://deepmind.google/technologies/gemini

[13] Ö. Aydin and E. Karaarslan, *OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare*, vol. 2. İzmir Akademi Dernegi, 2022. [Online]. Available: https://ssrn.com/abstract=4308687

[14] E. A. Oliveira, S. Rios, and Z. Jiang, "AI-powered peer review process: An approach to enhance computer science students' engagement with code review in industry-based subjects," in *ASCILITE 2023 Conference Proceedings: People, Partnerships and Pedagogies*, Christchurch, New Zealand, 2023.

[15] W. Dai *et al.*, "Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, IEEE, 2023, pp. 323–325. [Online]. Available: https://chat.openai.com/

[16] A. Matin *et al.*, "Trust in Generative AI among students: An Exploratory Study," in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[17] M. Amoozadeh *et al.*, “Towards Characterizing Trust in Generative Artificial Intelligence among Students,” *ICER ’23: Proceedings of the 2023 ACM Conference on International Computing Education Research*, vol. 2, Aug. 2023, doi: 10.1145/3617367.

[18] V. , Charisi *et al.*, “Artificial Intelligence and the Rights of the Child Towards an Integrated Agenda for Research and Policy,” Luxembourg, 2022. doi: 10.2760/012329.

[19] “AI Act: a step closer to the first rules on Artificial Intelligence,” European Parliament. Accessed: Jan. 23, 2024. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

[20] S. Hashim, M. K. Omar, H. Ab Jalil, and N. Mohd Sharef, “Trends on Technologies and Artificial Intelligence in Education for Personalized Learning: Systematic Literature Review,” *International Journal of Academic Research in Progressive Education and Development*, vol. 11, no. 1, Feb. 2022, doi: 10.6007/ijarped/v11-i1/12230.

[21] C. Ling Thong, R. Butson, and L. WeiLee, “Understanding the impact of ChatGPT in education: Exploratory study on students’ attitudes, perception and ethics,” in *ASCILITE 2023*, 2023. [Online]. Available: https://www.aleks.com

[22] A. Barredo Arrieta *et al.*, “Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, no. December 2019, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[23] D. Gunning, “Explainable artificial intelligence (xai),” *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, no. 2, 2017, [Online]. Available: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning) IJCAI-16 DLAI WS.pdf

[24] H. Khosravi *et al.*, “Explainable Artificial Intelligence in education,” *Computers and Education: Artificial Intelligence*, vol. 3, no. May, 2022, doi: 10.1016/j.caeai.2022.100074.

[25] L. Alrajhi, F. D. Pereira, A. I. Cristea, and T. Aljohani, “A Good Classifier is Not Enough: A XAI Approach for Urgent Instructor-Intervention Models in MOOCs,” in *Artificial Intelligence in Education (AIED 2022)*, 2022, pp. 424–427. doi: 10.1007/978-3-031-11647-6_84.

[26] T. Atapattu, K. Falkner, M. Thilakaratne, L. Sivaneasharajah, and R. Jayashanka, “What Do Linguistic Expressions Tell Us about Learners’ Confusion? A Domain-Independent Analysis in MOOCs,” *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 878–888, Oct. 2020, doi: 10.1109/TLT.2020.3027661.

[27] H. Du and W. Xing, “Leveraging explainability for discussion forum classification: Using confusion detection as an example,” *Distance Education*, vol. 44, no. 1, pp. 190–205, 2023, doi: 10.1080/01587919.2022.2150145.

[28] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke, “YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips,” *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 297–304, 2015, [Online]. Available: http://ilpubs.stanford.edu:8090/1125/1/you_edu.pdf

[29] O. Almatrafi, A. Johri, and H. Rangwala, “Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums,” *Comput Educ*, vol. 118, pp. 1–9, Mar. 2018, doi: 10.1016/J.COMPEDU.2017.11.002.

[30] A. Bakharia, “Towards cross-domain MOOC forum post classification,” in *L@S 2016 - Proceedings of the 3rd 2016 ACM Conference on Learning at Scale*, Association for Computing Machinery, Inc, Apr. 2016, pp. 253–256. doi: 10.1145/2876034.2893427.

[31] A. Chanaa and N. E. El Faddouli, “BERT and Prerequisite Based Ontology for Predicting Learner’s Confusion in MOOCs Discussion Forums,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2020, pp. 54–58. doi: 10.1007/978-3-030-52240-7_10.

[32] Z. Zeng, S. Chaturvedi, and S. Bhat, “Learner Affect Through the Looking Glass: Characterization and Detection of Confusion in Online Courses,” in *the 10th International Conference on Educational Data Mining*, 2017, pp. 272–277.

[33] G. Bhumireddy and V. A. S. M. Anala, “Comparison of Machine Learning algorithms on detecting the confusion of students while watching MOOCs,” Master of Science, Blekinge Institute of Technology, Karlskrona, Sweden, 2022. Accessed: Dec. 14, 2022. [Online]. Available: www.bth.se

[34] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[35] M. Jojoa, P. Eftekhar, B. Nowrouzi-Kia, and B. Garcia-Zapirain, “Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization,” *AI Soc*, 2022, doi: 10.1007/s00146-022-01594-w.

[36] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[37] N. Kokhlikyan *et al.*, “Captum: A unified and generic model interpretability library for PyTorch,” pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2009.07896

[38] Y. Hu, R. F. Mello, and D. Gašević, “Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence,” *Computers and Education: Artificial Intelligence*, vol. 2, p. 100037, 2021, doi: 10.1016/j.caeai.2021.100037.

[39] C. Li and W. Xing, “Natural Language Generation Using Deep Learning to Support MOOC Learners,” *Int J Artif Intell Educ*, vol. 31, no. 2, pp. 186–214, Jun. 2021, doi: 10.1007/S40593-020-00235-X/FIGURES/6.

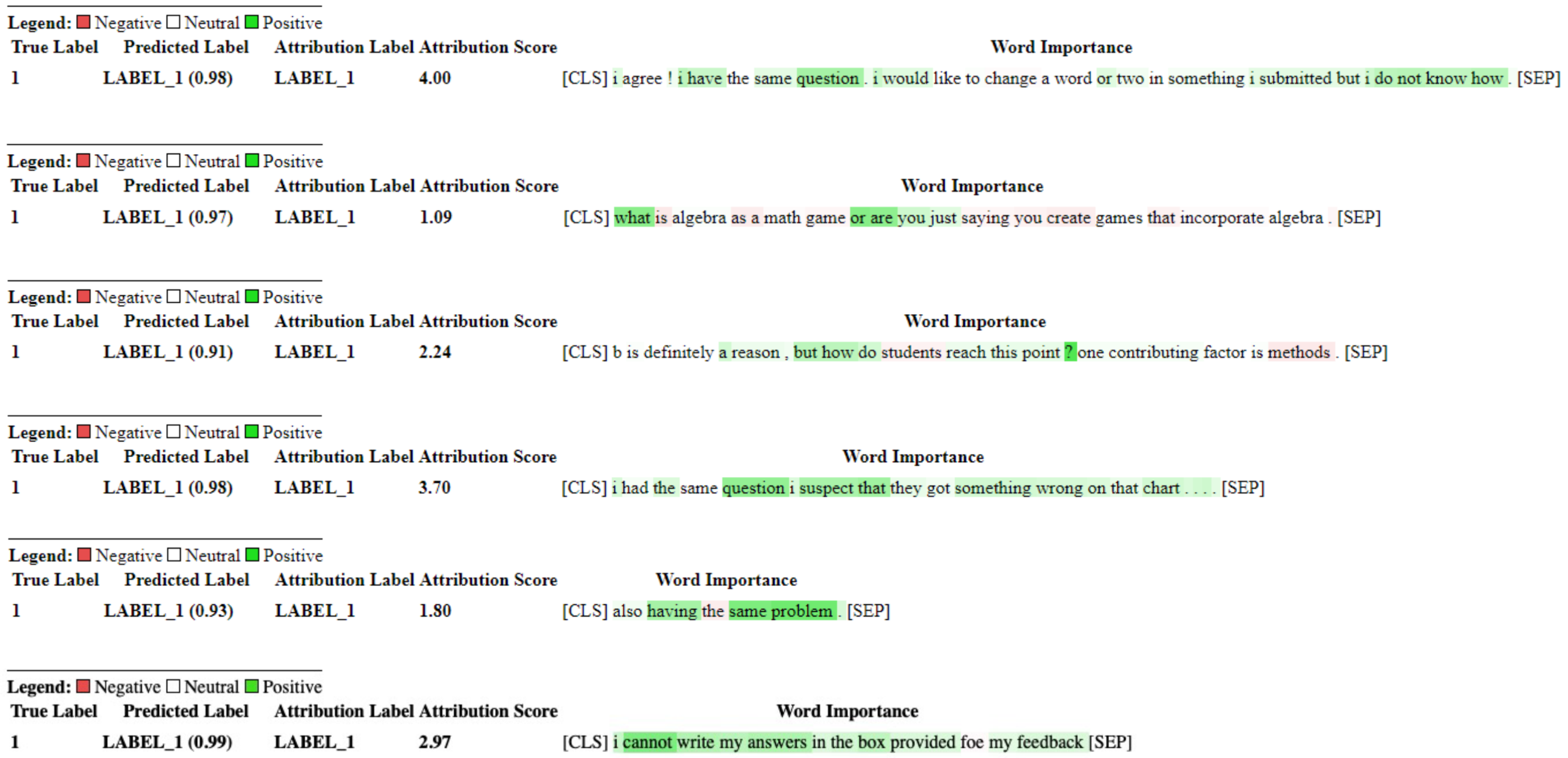
[40] S. Knight *et al.*, “AcaWriter A learning analytics tool for formative feedback on academic writing,” *J Writ Res*, vol. 12, no. 1, pp. 141–186, 2020, doi: 10.17239/JOWR-2020.12.01.06.

[41] D. Clow, “The learning analytics cycle: Closing the loop effectively,” in *the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12*, Vancouver, BC., 2012, pp. 134–138. doi: 10.1145/2330601.2330636.

[42] L. Yan, R. Martinez-Maldonado, and D. Gašević, “Generative Artificial Intelligence in Learning Analytics: Contextualising Opportunities and Challenges through the Learning Analytics Cycle,” Nov. 2023. [Online]. Available: http://arxiv.org/abs/2312.00087

[43] M. Sharples, “Towards social generative AI for education: theory, practices and ethics,” *Learning: Research and Practice*, vol. 9, no. 2, pp. 159–167, Jun. 2023, doi: 10.1080/23735082.2023.2261131.

# Appendix



**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.98) | LABEL_1 | 4.00 | [CLS] i agree ! i have the same question . i would like to change a word or two in something i submitted but i do not know how . [SEP] |

**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.97) | LABEL_1 | 1.09 | [CLS] what is algebra as a math game or are you just saying you create games that incorporate algebra . [SEP] |

**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.91) | LABEL_1 | 2.24 | [CLS] b is definitely a reason , but how do students reach this point ? one contributing factor is methods . [SEP] |

**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.98) | LABEL_1 | 3.70 | [CLS] i had the same question i suspect that they got something wrong on that chart . . . . [SEP] |

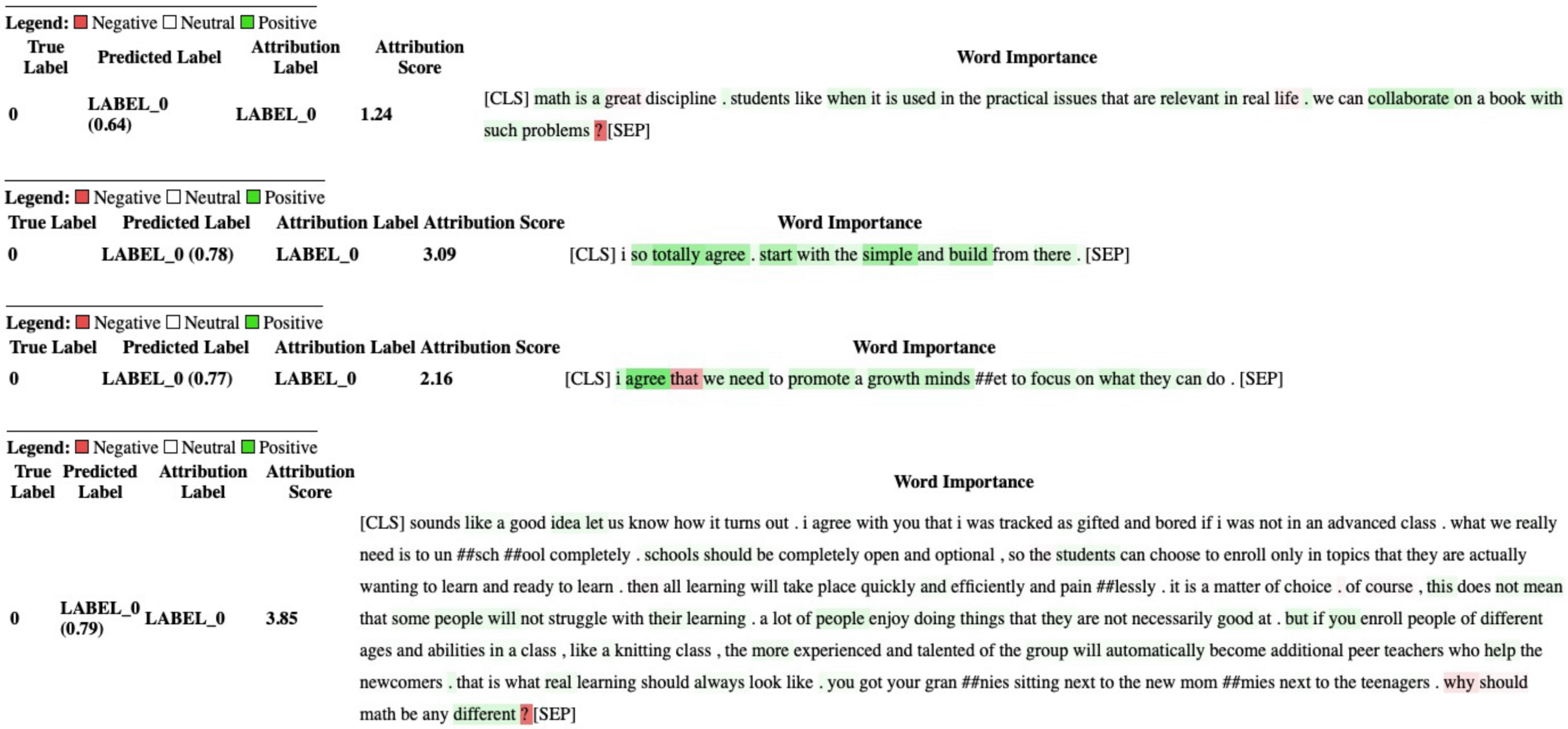
**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.93) | LABEL_1 | 1.80 | [CLS] also having the same problem . [SEP] |

**Legend:** 🟥 Negative ☐ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 1 | LABEL_1 (0.99) | LABEL_1 | 2.97 | [CLS] i cannot write my answers in the box provided foe my feedback [SEP] |

**Figure 1**: Samples of confusion messages in the Education Dataset



Legend: 🟥 Negative □ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 0 | LABEL_0 (0.64) | LABEL_0 | 1.24 | [CLS] math is a great discipline . students like when it is used in the practical issues that are relevant in real life . we can collaborate on a book with such problems ? [SEP] |

Legend: 🟥 Negative □ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 0 | LABEL_0 (0.78) | LABEL_0 | 3.09 | [CLS] i so totally agree . start with the simple and build from there . [SEP] |

Legend: 🟥 Negative □ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 0 | LABEL_0 (0.77) | LABEL_0 | 2.16 | [CLS] i agree that we need to promote a growth minds ##et to focus on what they can do . [SEP] |

Legend: 🟥 Negative □ Neutral 🟩 Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| 0 | LABEL_0 (0.79) | LABEL_0 | 3.85 | [CLS] sounds like a good idea let us know how it turns out . i agree with you that i was tracked as gifted and bored if i was not in an advanced class . what we really need is to un ##sch ##ool completely . schools should be completely open and optional , so the students can choose to enroll only in topics that they are actually wanting to learn and ready to learn . then all learning will take place quickly and efficiently and pain ##lessly . it is a matter of choice . of course , this does not mean that some people will not struggle with their learning . a lot of people enjoy doing things that they are not necessarily good at . but if you enroll people of different ages and abilities in a class , like a knitting class , the more experienced and talented of the group will automatically become additional peer teachers who help the newcomers . that is what real learning should always look like . you got your gran ##nies sitting next to the new mom ##mies next to the teenagers . why should math be any different ? [SEP] |

**Figure 2**: Samples of non-confusion messages in the Education Data set