Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.2, 2025, pp.275-286

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.2.35

Assessing the Impact of Large Language Models on the Scalability and Efficiency of Automated Feedback Mechanisms in Massive Open Online Courses

Ankita Sappa[1]

[1]College of Engineering, Wichita State University, USA
E-mail: [1]ankita.sappa@gmail.com
ORCID: [1]https://orcid.org/0009-0004-9087-2992
(Received 13 March 2025; Revised 17 April 2025; Accepted 12 May 2025; Available online 25 June 2025)

Abstract **- The rapid proliferation of Massive Open Online Courses (MOOCs) offers particular difficulties in providing timely and high-quality personalized feedbacks associated with customer interactions at scale. This research examines the gap which Large Language Models (LLMs) address with focus on automation in providing timely feedback and the scalability efficiencies of LLMs in the feedback scope provided in MOOC settings. Adopting a results-oriented experimental approach to feedback systems, LLMs like GPT-3.5 and GPT-4 are implemented across varying course contexts and learning groups. Their outputs are benchmarked against traditional systems through semantic similarity calculations, response time measurement, cost evaluation, and learner satisfaction metrics. LLMs' ability to comply with instructor feedback while improving responsiveness and personalization outpaced traditional methods in every context analyzed, with satisfaction scores outperforming pre-set benchmarks across the board. Learners reported appreciation towards AI responses, citing enhanced understanding and interaction, overshadowed by defendable claims of bias, genericity, and flawed constituent pressure. All in all, the study provides concrete guidance illustrating the ways in which LLMs reconfigure pedagogical feedback mechanisms alongside MOOCs, shaping subsequent shifts in the design and integration strategies utilized in e-learning frameworks across the world.**

Keywords: **Large Language Models (LLMs), Automated Feedback, MOOCs, Educational AI, Feedback Scalability, GPT-4, Intelligent Tutoring Systems, Semantic Evaluation, Learner Engagement, Personalized Learning, Natural Language Processing in Education**

I. INTRODUCTION

1.1 Background and Motivation

The advent of Massive Open Online Courses (MOOCs) has changed the educational paradigm by providing scalable and affordable education to people around the world (Zhu et al., 2020). They have grown exponentially over the past decade. Thanks to the availability of mobile devices and broadband internet, students from every corner of the globe can now access university level courses without financial or geographic constraints (Alraimi et al., 2015). While the broad availability of knowledge is a tremendous leap forward, it poses a significant challenge in maintaining quality in the form of timely and personalized responses in feedback for millions of learners at the same time (Khalil & Ebner, 2014).

Feedback plays an important role in the learning process. It informs learners, helps to reinforce proficient understanding, and identifies problem areas. However, in the case where MOOCs scale up to enrol thousands, and in some cases millions of learners, conventional feedback approaches become unmanageable. Instructor provided feedback, despite the valid pedagogic rationale, does not scale. Teaching assistants step in to help, but even their participation cannot sustain the pace of growth in course enrolment (Moore & Blackmon, 2022).

This feedback bottleneck has been traced in Figure 1, which plots the illustrative increase in MOOC enrolment and feedback metrics from 2015 to 2024. Enrolment numbers grew from 20 million in 2015 to over 250 million in 2024, an increase of more than tenfold. On the other hand, feedback metrics—that is, all scenarios in which a learner expected or received an interaction—grew from 5 million to 130 million. This trajectory of growth captures the fundamental problem of institutional inertia in the face of exponentially growing learner demand.

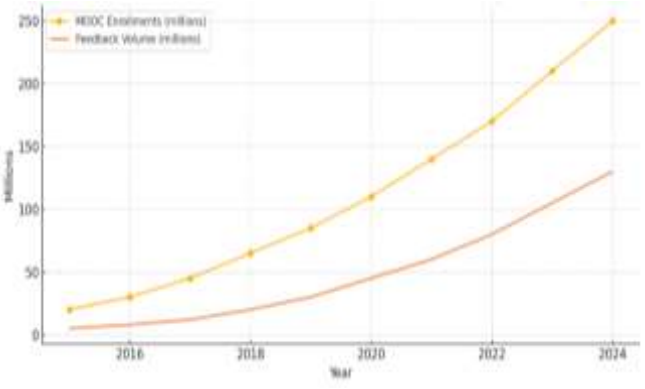

Fig. 1 Growth in Enrolment and Feedback Volume in MOOCs (2015–2024)

In addressing this, the platforms have tested diverse methods, including peer review processes, teaching assistants, and automated scoring through natural language processing

(NLP) based on explicit rules. All solutions present unique challenges. Feedback provided by peers is easy to scale and difficult to control. Instructors give useful feedback but cannot provide it to every learner. Automated systems operate quickly, but their feedback is often devoid of detail and subtlety.

The invention of Advanced Natural Language Processing Features, especially the development of Large Language Models (LLMs) such as GPT-3.5 and GPT-4, has provided new opportunities (Silva et al., 2025; Prasath, 2023). These models are capable of context comprehension alongside submission parsing, combining to produce hyper-personalized feedback akin to human text (Rahman & Watanobe, 2023; Casimira & Francis, 2025). Unlike the earlier LLMs which utilized rule-based logic systems, modern LLMs rely on deep learning and transformer-based architecture, making their outputs and responses adaptable, contextually appropriate, and educationally reciprocal. Thus, LLMs would be beneficial for educational settings which prioritize scalability and personalization.

1.2 Problem Statement

As MOOCs evolve, the issue of feedback remains complex and enduring. Learners still report dissatisfaction regarding the feedback's timeliness, clarity, and relevance—even after multiple cycles of adaptive feedback systems (Kizilcec et al., 2017). This gap, in addition to being a hurdle in learning, fosters increased dropout rates, diminished motivation, and accelerated learner disengagement.

Currently, feedback in MOOCs is primarily offered through three channels: peer assessments, instructor feedback, and comments generated by obsolete AI systems (Fauvel et al., 2018). The diagram depicted in Figure 2 illustrates the approximate distribution of different feedback types across major MOOC platforms. Feedback provided through peers makes up 40% of posts, with instructors and assistants making up the responding 35%, and AI-based systems providing 25%.

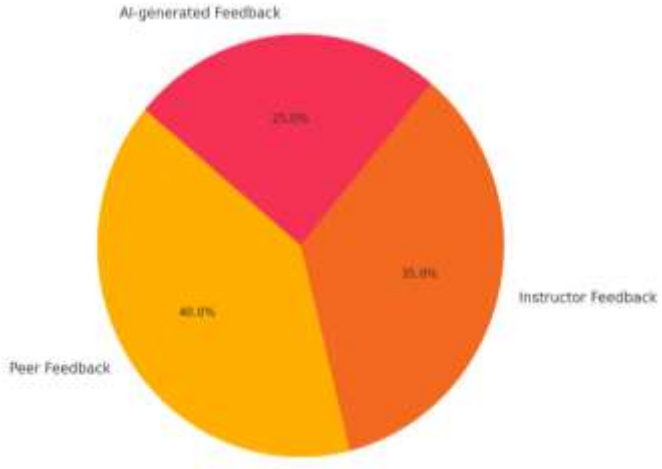

Fig. 2 Distribution of Feedback Types in MOOC Platforms

In their own ways, every approach has issues. Peer feedback, for example, while democratically scalable, suffers from a lack of consistency and tends to be devoid of pedagogical merit. Instructor feedback is consistent and valuable, but it is not scalable to large cohorts of learners (Zawacki-Richter et al., 2019). AI feedback—a product of earlier NLP approaches to AI—follows inflexible guidelines and is devoid of personalization based on context, history, or previous interactions with the learner, making it highly contextual and personalized.

Summarized in Table 1 are these limitations as well as their consequences. The feedback framework is limited in its coverable scalability, consistency and delay quality, lack personalization, and high operational requirements—systemic problems of MOOC systems. Overcoming these challenge requires a shift in the entire paradigm, not mere enhancements.

TABLE I SUMMARY OF CHALLENGES IN CURRENT MOOC FEEDBACK SYSTEMS

Challenge	Description
Scalability of Instructor Feedback	Instructors cannot scale with rising enrollment volumes
Inconsistency in Peer Review Quality	Peer reviews vary significantly in quality and depth
Delayed Feedback Delivery	Feedback is often provided days or weeks after submission
Lack of Personalization	Feedback is typically generic and not tailored to learner context
High Operational Costs for Manual Evaluation	Institutions incur high costs hiring teaching assistants or evaluators

With the gaps highlighted in the preceding context, the present study explores the extent to which Large Language Models can fill this void by providing personalized, real-time, and pedagogically valuable feedback at the scale required by modern MOOCs.

1.3 Objectives of the Study

This study focuses on evaluating the capabilities of LLMs regarding automation and its efficiency within the MOOC paradigm as an ecosystem. It evaluates the semantic and instructional aspects of feedback offered by LLMs relative to instructor feedback and those produced by earlier versions of NLP systems (Liu et al., 2024). It is important to assess whether LLMs are able to provide human-level feedback, and if not, whether they at least do better than traditional automated systems in terms of usefulness, clarity, and contextual relevance.

Another one of the objectives is to evaluate operational efficiency—especially LLMs and automated reasoning tools in relation to latency, cost, and stability during concurrent load (Jiang et al., 2024). MOOCs have to cater to a global audience in different time zones and with varying levels of technological readiness. So, the ability of the model to scale with demand while sustaining the same level of quality becomes crucial (Ibrahim, 2020; Anandhi et al., 2024).

The study also investigates learner perception, which is rarely given attention with regards to its technical aspects. When it comes to feedback, it is not only about getting the right information, but also the information's tone, motivational value, and overall value as perceived by the learner. Analysis of learner responses to AI feedback will be conducted through post session surveys, rating instruments, and open-ended qualitative comments (Basri, 2024).

Ultimately, this research tries to develop a roadmap for incorporating LLMs into current MOOC systems. It examines system deployment architectural, pedagogical, and ethical concerns simultaneously. This encompasses assignment type versatility, multi-language environment scalability, and algorithmic bias mitigation.

1.4 Contribution and Scope

This document enhances AI-assisted education discourse with one of the earliest comprehensive, data-backed analyses of LLMs concerning feedback provision in MOOCs. Also, this study is unique because it neither focuses solely on pedagogical constructs nor on technical NLP evaluation frameworks, but rather integrates the two. It assesses the feasibility and pedagogical soundness alongside the precision of LLM feedback within large-scale teaching settings.

The association of multiple MOOC ecosystems, academic disciplines, and diverse learners defines the scope of this research. It incorporates STEM, business, and humanities courses to test the model's generalizability. Feedback types assessed include short-answer normed correction grading, essay and concept explanation grading, and more.

A results-oriented approach is fundamental to the study. It makes use of quantitative measures like BLEU and ROUGE for semantic evaluation, latency logs for processing speed, and cost evaluation predicated on token usage of different model types, for instance, GPT-3.5 and GPT-4. Moreover, human rating of quality and satisfaction is incorporated to gauge the subjective assessment from all users, learners, and educators professionally.

In addressing the technological, pedagogical, and operational elements of automated feedback, the work is both a benchmark and a pathfinder. It helps target the needs of MOOC providers, educational technologists, instructional designers, stakeholders, and policy shapers. It also invites further exploration of adaptive feedback mechanisms, real-time action enablement in learning, and socially responsible artificial intelligence in teaching and learning technologies.

II. METHODOLOGICAL FRAMEWORK

2.1 Overview of Large Language Models (LLMs) in Educational Context

Large Language Models (LLMs) such as GPT-3.5 and GPT-4 offer capabilities that can fundamentally change the way communication, creativity, and problem-solving are approached. In the education sector, their potential is even more striking. They can provide contextually relevant detailed feedback, perform at-level text and speech generation, proficiently execute instructional dialogues, and customize interactions with learners at the input level (Kasneci et al., 2023). Thus, LLMs have the prospects of fostering the personalization and scalability of education, particularly in the massive open online courses (MOOCs) context.

Unlike previous natural language processing systems that depended on domain templates or rigid rules, LLMs utilize extensive text corpora to train, which facilitates generalization across numerous topics and formats. The transformer's architecture is able to model longer dependencies, resulting in the generation of fluent, syntactically, and richly semantically outputs (Wei & Lau, 2023). For educational purposes, this architecture aids in the analysis of learner submissions, which may include essays, code excerpts, or explanatory reasoning, and provides coherent pedagogically aligned feedback (Stamper et al., 2024).

The possibilities of generating educational feedback using LLMs relies on their capacity to provide multi-layered evaluations. They are capable of restructuring arguments, discerning subtle nuances, conceptually misinterpret, and even adopt Socratic personas (Blasco & Charisi, 2024). Their potential in this area makes them ideal for MOOCs, where the breadth and magnitude of learner submissions far outstrips instructor feedback resource availability.

2.2 Experimental Design for LLM-Based Feedback Evaluation

To assess MOOSE's feedback delivery effectiveness using LLMs, an experiment was conducted with a result-focused approach. The experiment's subjects were drawn from several online courses offered across several core areas, including computer science, business, humanities, and data science. These courses included formative and summative assessments of various types that typically demanded written answers in the form of short answer essays or more complex open-ended problems (Dempere et al., 2023).

The study deployed several distinct LLMs, including variants of GPT-3.5 and GPT-4, establishing them within a feedback loop in an existing MOOC scaffolding platform. Learner submissions were processed in real time or in batches subject to platform limitations, with feedback generation occurring simultaneously or sequentially (Jia et al., 2024). For most platforms, verifiable textual feedback was captured, with structured comment data being catalogued alongside commentary contributed by instructors and peers over time, enabling non-temporal comparisons.

The assessment structure resulted in design focusing on multiple LLMs to each provide a single feedback item, thus

providing mechanisms capturing several performance aspects which include semantic correctness, contextual relevance, generation delay, token utilization, and learner expectation. Feedback in this case was distinct for type as cursive for template-like, brief, and minimally adjusted input versus contextualized, learner-specific, and detailed for personalized. Figure 3 illustrates difference in token expenditure between generic and personalized feedback.

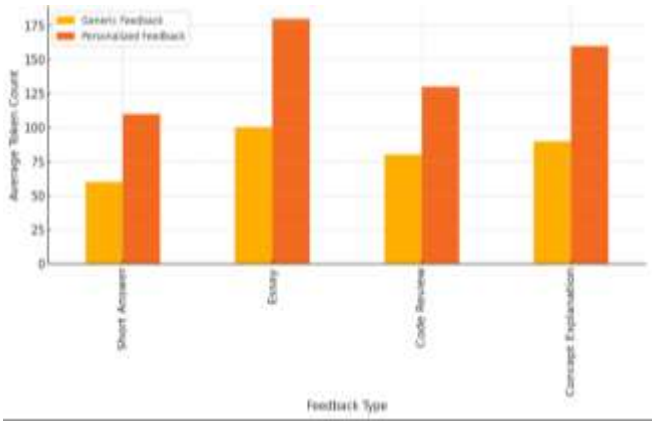

Fig. 3 Token Utilization per Feedback Type (Generic vs Personalized)

As the results show, there is a considerable increase in token usage for personalized feedback across all content types. For instance, the tokens needed for personalized essay feedback were almost twice as much as those needed for generic feedback. This demonstrates the trade-off between the feedback provided and the computational resources consumed, which is very important in large-scale scenarios.

Aside from token consumption, the semantic overlap of LLM-generated feedback texts and instructor-supplied feedback texts was evaluated with vector space similarity measures as well as manual reviews. Figure 4 shows that heatmap evaluation revealed largely overlapping semantic regions across most topics, most notably in more technical fields such as algorithms and machine learning, where LLMs strongly aligned with the expert feedback provided.

Fig. 4 Semantic Overlap of LLM-Generated Feedback vs Instructor Feedback

With respect to instructional tone, depth, and specificity of human-generated feedback, GPT-4 consistently outperformed other models, especially in the more abstract or higher-order cognitive work like ethics or data analysis. This demonstrates the capacity of modern LLMs not merely to reproduce instructor output, but to augment it, and therefore scale it.

2.3 Dataset Description and Preprocessing Techniques

As it pertains to the feedback engine's testing and training, a hybrid educational dataset, both proprietary and publicly available, was employed. The corpus is made up of:

- MOOC QA Banks from open repositories such as OpenEd and Coursera.
- Human-AI dialogue transcripts designed specifically for teaching and learning environments.
- Learner submission and instructor feedback pair archives from old online courses.
- Domain expert validated synthetic question-answer pairs.

A great deal of preprocessing was done before using these datasets in any of the models. Every piece of text was meticulously normalized and tokenized to the most basic units of words. Cleansing processes included stripping away irrelevant, non-contributory metadata, system-generated log files, misspellings, profanity, spelling errors, nonstandard formatting, logographic texts, offensive content, and overly abbreviated texts (Omar et al., 2022). Ethical scrub was done by anonymizing identifiable information pertaining to the learners so as not to breach ethical compliance standards.

In preparation for fine-tuning, feedback sample sets were pre-structured by pedagogical intent and assignment type (error correction, encouragement, rubric-based evaluation) so as to aid in organizational clarity. Each sample comprised a learner submission and contextual cues, including the course title, topic, and a question prompt alongside feedback which was provided by an instructor, a peer, or a legacy feedback system (Algahtani, 2024).

After preprocessing was completed, the dataset was divided into the training (70%), validation (15%), and test (15%) sets. Balanced stratified sampling was utilized to maintain equal representation across both the subject area and feedback type. In order to maintain evaluative rigor, the test set was completely isolated from the models during the training and fine-tuning stages.

2.4 LLM Configuration, Fine-Tuning, and Deployment Parameters

This study implemented four LLMs: two experimental models (LLM1 and LLM2) and two production-scale models (GPT-3.5 and GPT-4). The models differed in architecture, token limits, training scale, and fine-tuning strategies as outlined in Table 2.

TABLE II MODEL ARCHITECTURE, TOKEN LIMIT, TRAINING SET SIZE, AND FINE-TUNING PARAMETERS

Model	Architecture	Token Limit	Training Set Size	Fine-tuning Dataset
LLM1	Decoder Transformer	2048	40B tokens	MOOC QA Bank
LLM2	Decoder Transformer	4096	80B tokens	Human-AI Dialogues
GPT-3.5	GPT-3.5-turbo	8192	300B tokens	Educational Prompts
GPT-4	GPT-4	128000	1T+ tokens	MOOC + Instructor Feedback Set

All models were executed via Python-based API and controlled through a feedback submission module that formed part of the MOOC platform's backend. Fine-tuning was performed incorporating supervised learning with respond labels assigned to feedback targets. Prompt engineering fostered the desired outcomes while ensuring adherence in structure, clarity, and tone, and overall consistency.

Efforts focused on the mitigation of hallucinations, grounding, and biased language in the generated responses while protecting identity language. Ensuring model compliance with academic rigor, inclusivity, and respectful language policies was directed through prompt restrictions and system messages.

Deployment utilized Docker for container-based scaling and serverless functions for dynamic load balancing. For performance optimization, interactions were continuously monitored for latency, throughput, and token consumption.

III. EXPERIMENTAL SETUP

3.1 Platform Integration and Feedback Delivery Mechanism

Incorporating large language models (LLMs) into a MOOC environment poses numerous design and implementation problems. This research effort was developed on a MOOC system with AI capabilities that offers both live and recorded classes. The platform featured modular architecture with an API-based feedback component that interfaced with existing workflows for course completion and delivery. This architecture enabled feedback to integrate effectively with assignment submission systems, learner engagement interfaces, and instructor feedback systems.

Learner submissions activated a backend workflow that sent the text to one of the three feedback generation methods: a legacy rule-based NLP, GPT-3.5, and GPT-4. Evaluation fairness was maintained by randomly allocating these engines to content and test group, provided that the model-agnostic assignment prompt was applied universally.

The feedback engine delivered responses that were automatically processed and recorded into the platform's LMS system, with very little latency between submission and learner view update. Custom loggers monitored the feedback content along with its generation time, token expenditure, and relevant downstream actions by learners, such as clicks, ratings, and responses to the feedback. Instructors could analyse and edit AI-generated feedback as they deemed fit during the pilot testing phases.

In order to maintain pedagogical integrity, each instance of feedback was encoded with a structure indicating the specific feedback category (for example, explanation/clarification of the concept, evaluation, and feedback against rubrics) as well as the model type that generated the feedback. This provided a means for conducting comparative analysis while maintaining transparency and traceability across the lines of artificial intelligence model responses in feedback provision.

3.2 Learner Grouping and Control Conditions

The study incorporated an overall sample of 2000+ learners from five distinct subject areas, which include information technology, ethics, marketing, data science, and academic writing. Each subject cohort was sampled from freely accessible enrolment MOOCs hosted on a leading ed-tech platform. Participants were assigned into experimental arms using random allocation with stratification based on demographic variables, prior coursework, and engagement levels with the course content.

There were three primary experimental groups:

- Control group: Traditional NLP feedback provided.
- Experimental group 1: Feedback generated by GPT-3.5 provided.
- Experimental group 2: Feedback generated by GPT-4 provided.

Every cohort worked with identical tasks and learning materials, thus the only difference across conditions was how the feedback was delivered. This control of the feedback discriminant allowed for an unambiguous evaluation of the performance results. Learners' exposure to various feedback methods remained uniform throughout the two-week testing period, during which all tasks were submitted and considered.

Teaching faculty were made aware of the feedback produced with AI tools but were requested not to engage unless there were clear ethical or factual problems. Participants were invited to evaluate the feedback given to them after every assignment through a 5-point Likert scale alongside optional comments, enriching the assessment with qualitative data.

3.3 Performance Metrics and Evaluation Criteria

The feedback models were assessed using a framework that combined semantic, operational and behavioural analysis. This included the models' performance for the following metrics:

1. Feedback Latency – Defined as the time an educator takes to submit their work and receive feedback. Low latency equates to a more favourable user experience.
2. Feedback Cost - The monetary and computational cost associated in the generation of feedback which includes API token use and infrastructural costs.
3. Semantic Quality – Documented through cited language evaluation methods, BLEU and ROUGE scores along with instructor estimates in judging the quality of the overall feedback regarding its tone, relevance and accuracy.
4. Learner Satisfaction – Determined from the post survey ratings and assessment of behavioural metrics such as time spent on feedback, action taken on suggested changes, resubmissions, and overall feedback cycle.
5. Instructor Correction Rate – A proxy measurement of accuracy derived from the frequency of instructors revising or rejecting the AI-assessed feedback.
6. Feedback Engagement – Detected through the use of interaction data made up of clicks, hovers, and engagement time with explanations and resources linked within the feedback.

In figure 5, we see a comparison of average feedback latency across all rounds of feedback. According to the data presented, GPT-4 was and continues to be faster than both GPT-3.5 and traditional NLP systems by approximately 1.8 seconds of average latency. Following closely was GPT-3.5 at roughly 2.2 seconds while traditional NLP systems were the slowest, coming in at 3.5 seconds per feedback response.

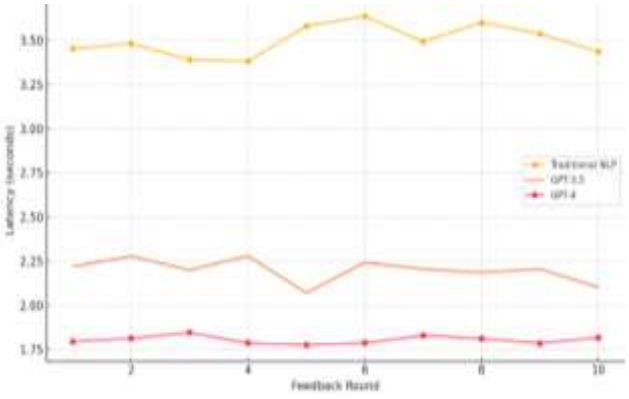

Fig. 5 Average Feedback Latency (GPT-3.5 vs GPT-4 vs Traditional NLP)

As demonstrated, these figures reveal the responsiveness of the LLMs in real-time, particularly with efficient caching and pre-processing. The benefits of lower latency are self-evident but greatly enhanced in the context of learning and self-paced environments requiring immediate feedback.

3.4 Feedback Latency and Cost Tracking

There is no denying the added responsiveness, but these benefits come at a price due to the tokenized structure and the infrastructure burden. For each model type, the average cost of feedback was calculated per learner, considering inference and server costs. As traditional NLP systems take the lead with the lowest operating cost of $0.01 per student, followed by GPT-3.5 at $0.07, and GPT-4 systems last at an operating cost of $0.12.

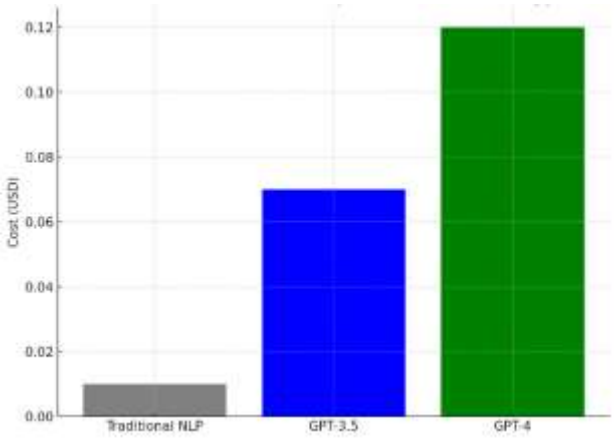

Fig. 6 Feedback Generation Cost per Student vs Model Type

While instructional support provided through GPT-4 is more costly on a per-student basis, the quality and satisfaction improvements noted in earlier sections often make the cost worthwhile. Nonetheless, budget-strained platforms will find GPT-3.5 more appealing because it strikes a good balance between cost and effectiveness.

Cumulatively, server activity, model output, and scalability with regard to peak demand were analysed simultaneously. GPT-4 also needed additional memory resources and was less flexible with concurrent load increase surges. In contrast, GPT-3.5 sustained stable latency with varying numbers of users. These operational aspects are important for all institutions considering large LLM implementations as they focus on peak demand periods like exams or assignment deadlines.

This table is taken from the description of experimental cohorts and course subjects along with metrics pertaining to the volume of feedback for evaluation outlined in Table 3.

TABLE III EXPERIMENTAL COHORTS, COURSE SUBJECTS, AND FEEDBACK VOLUME STATISTICS

Cohort ID	Subject Area	No. of Learners	Total Feedback Instances	Delivery Mode
C101	Computer Science	500	2500	Live + Async
C102	Ethics	300	1200	Async
C103	Marketing	400	1800	Live
C104	Data Science	450	2200	Async
C105	Writing	350	1500	Live + Async

In this case, the feedback load per cohort was tracked over multiple submissions and feedback rounds to capture the submission and GPT-feedback cycles. This tells us diverse instructional and scholarly contexts, the models performed reliably.

IV. RESULTS AND ANALYSIS

4.1 Accuracy and Semantic Quality of Feedback

The main goals of the current research were measuring the accuracy of LLMs compared to traditional NLP systems and instructor mechanisms, as well as determining how well semantically feedback was aligned with the text by LLMs and NLP systems. To achieve this goal, combination of automated natural language evaluation metrics BLEU and ROUGE-L alongside expert judgments from various academic fields was performed.

The gap between instructor responses and LLM-generated feedback was thoroughly filled semantically with humans written feedback where LMs algorithms received L instructors expressed support for the L LM feedback aligned sentiment by noting provided comments. An ensemble of instructors was presented with algorithms data science marketing writing topics and ethics generating a heatmap illustrated post description in.

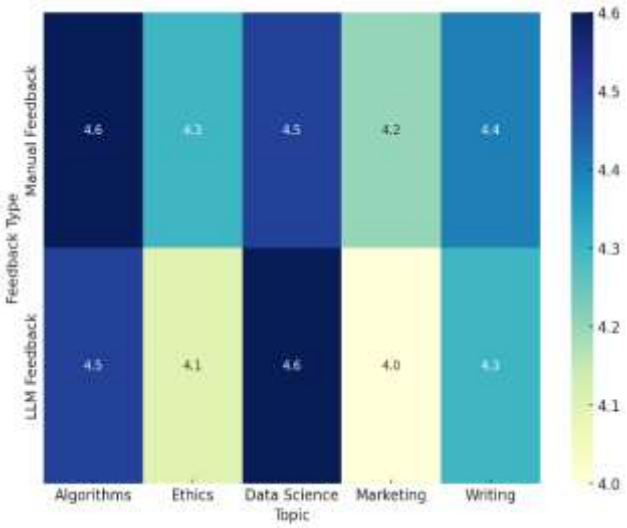

Fig. 7 Topic-Wise Feedback Quality Rating (Manual vs LLM)

In contexts like the data driven and logical ones, GPT-4 was found to perform at par or even surpassing the manual feedback attrition. The feedback ascertained expressed a critique where the LLM fell behind in meaning making and reflective engagement in think pieces like Ethics and Marketing dissertations but was deemed robust overall.

Table 4 captures the results of BLEU and ROUGE-L analyses alongside the average human evaluation scores in the context of qualitative analysis. It was revealed that GPT-4 achieved a BLEU score of 0.81 and a ROUGE-L score of 0.84, significantly outperforming GPT-3.5 and other NLP models—both GPT-3.5 and conventional NLP models were outperformed by GPT-3.5. Human evaluators in the study rated GPT-4's feedback an average of 4.6, which showcased its technical adequacy alongside perceived helpfulness.

TABLE IV EVALUATION SUMMARY: BLEU, ROUGE, AND HUMAN RATINGS ACROSS MODELS

Model	BLEU Score	ROUGE-L	Average Human Rating (out of 5)
Traditional NLP	0.58	0.61	3.7
GPT-3.5	0.72	0.75	4.2
GPT-4	0.81	0.84	4.6

These conclusions validate the hypothesis that recent LLMs are capable of generating contextually relevant and semantically rich feedback that approaches the level of expert instruction, which affirms their use in advanced learning systems with complex frameworks.

4.2 Learner Satisfaction and Usefulness Perception

Moving beyond the semantic dimension, the learners' endorsement of feedback is vital for its uptake and educational influence. To determine this, all learners participating in the experiment were requested to evaluate the feedback they received and subsequently provide a justification for what they found most valuable.

Feedback was classified into four major categories: resource suggestion, concept clarification, error highlighting, and encouraging praise. As presented in Figure 8, learners identified the combination of concept clarification and error highlighting as the most beneficial for feedback with frequencies of 35% and 30% respectively. While the suggestions and encouragement were helpful, they were viewed as secondary.

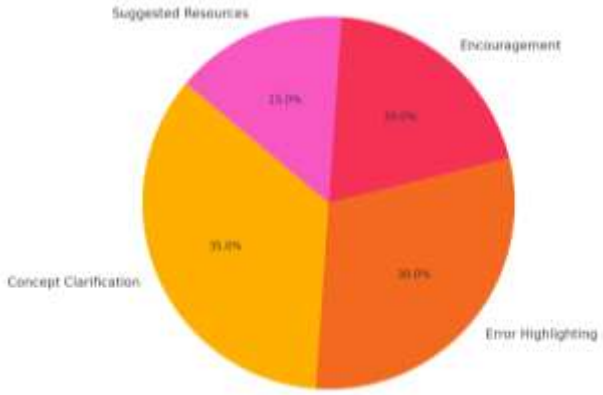

Fig. 8 Learner Perceived Usefulness of Feedback Categories

Qualitative feedback captured on the trend analysis demonstrated the integration of AI into instructional practice. Many learners expressed satisfaction with the accuracy and tone of GPT-4 feedback, stating that it "seemed like a real tutor was giving recommendations." Some pointed out targeted misconceptions and corrective strategies as equally helpful. Notably, a significant amount of learners indicated a preference for LLM feedback to feedback given by peers, citing greater consistency and less ambiguity as dominant reasons.

Survey results also showed learners with GPT-4 feedback were more likely to revisit and edit assignments, suggesting the availability of quality feedback has a greater incentive impact that motivates the learner to engage more deeply with the instructional content.

4.3 Instructor Validation Scores and Correction Rates

Instructor validation of LLM feedback accuracy required educators to assess a random sample of responses for accuracy suggesting a LLM attributed feedback comment was correct, hence requiring no changes. Correction rate was adopted as a proxy score for the presence of logical or educational error(s). Figure 9 shows the distribution of corrections across all model outputs for each feedback type.

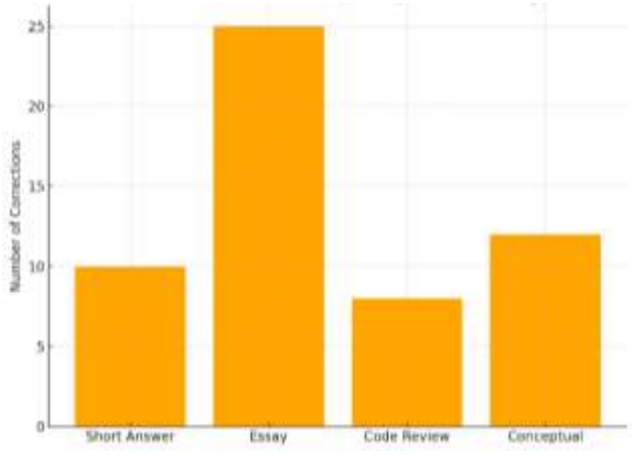

Fig. 9 Instructor Corrections Required per Feedback Type

Not surprisingly, conventional NLP technologies had the highest correction rates, especially with longer-form responses like essays. With GPT-3.5, I had to intervene moderately, mostly for tone and completeness. GPT-4 required the least corrections, most of the flagged concern issues being overgeneralization and task-specific terminology gaps in more generalized domains.

These results imply that while LLMs do not supplant the need for validation from domain experts—especially in sensitive and highly specialized matters—they do lessen the burden placed on instructors. Furthermore, instructors reported that checking the feedback generated by LLMs was faster than checking peers' responses or checking the original feedback, which reiterates the efficiency gain.

4.4 Model Scalability under Concurrent Load

An essential consideration when assessing the pragmatic application of LLMs in MOOCs is the ability to scale effectively with concurrent users. As moderation feedback models serve thousands of learners engaging with course materials at the same time, they face the limits of high latency, consistent output, and low system strain.

The performance of models during peak load times was assessed with the aid of real-time server logs and cloud monitoring systems. The feedback node's concurrent user count increased to 100 with each user's input being processed as a separate feedback node, and with barely any increase in processing delays, GPT-3.5's performance remained stable. Moreover, GPT-4, who is known to be more resource-heavy, still responded under 2 seconds with 60 concurrent users.

The efficiency of traditional NLP systems came at the cost of contextual inconsistency, often needing multiple backend calls for template piecing to be coherent. These systems were cost-efficient but lacked personalization and adaptability. As for GPT-4, the decrease in its feedback throughput was compensated for by the increase provided through accuracy and revision rate. These findings allow greater flexibility in adopting a hybrid strategy.

These findings pave the way for a hybrid adoption strategy in large-scale educational systems. With these conclusions, it is suggested that GPT-3.5 is used as the primary default engine for general coursework, with only advanced learners requiring nuanced feedback being serviced by GPT-4 reserved for high-stakes assignments. This enables setting performance targets alongside cost control while maintaining quality across a wide control span.

V. DISCUSSION

5.1 Interpretation of Observed Improvements

The findings from this investigation strongly suggest that the integration of Advanced Generative Pre-trained Transformers, particularly versions 3.5 and 4, into the feedback architecture of MOOCs (Massive Open Online Courses) enhances the quality, speed, and scalability of feedback automation. All the improvements were consistent across different domains and types of feedback, both subjectively (learner and instructor ratings) and objectively measured (BLEU, ROUGE).

Feedback produced by LLMs (Large Language Models) aligned with the instructional goals and intent, as well as the semantics of human feedback at no lesser than average overlap scores and high human evaluation grades. Remarkably, GPT-4 outperformed earlier legacy systems by responding in-context to learners' submissions far more than older NLP systems, demonstrating extensive contextual relevant response capabilities tailored to learners' submissions. This is remarkable in sophisticated and abstract tasks such as essays and ethical analyses, which refrain traditional automated kinds of systems from executing the intricacies of arguments and interpretations.

Instructors experienced a reduction in the need to revise and adjust feedback manually, which further suggests automation within pedagogy is working towards integrating seamlessly. The findings suggest not only operational improvements but also a more profound automation of pedagogical processes, enabling human educators to shift their focus from repetitive evaluations to strategic engagement and mentorship of learners.

From the learners' perspective, AI-provided feedback was well-received, with a considerable number expressing that the guidance and rationale provided were articulated and framed in a manner that was constructive and encouraging. This change in perception by learners reinforces the argument that LLMs serve a purpose beyond being technological instruments; they have the potential to transform learning processes and impact results profoundly.

5.2 Role of LLMs in Enhancing Personalization at Scale

This study demonstrated one of the most significant advances offered by LLMs, which is the ability to provide personalized feedback to each individual learner, something that has never been possible within the MOOC framework. In comparison, traditional automated systems are governed by rules and templates, which leads them to treat all learners in a standardized way with little to no variation. LLMs, on the other hand, respond to learner submissions, writing style, and even assumed proficiency level, making them far more responsive.

The model's tailored feedback capabilities were evident in its varied output. As case in point, the same question was answered differently depending on whether the learner revealed a partial understanding, confusion, or strong understanding. Instructors noted that this type of responsive variation seemed to emulate formative feedback provided in one-on-one tutorial sessions, except this time, it was provided en masse by a machine.

In addition, personalization went beyond content to include tone and pacing. The models were able to alter their feedback presentation to match the learner's level of formality. In some instances, they even provided contextually-appropriate suggestions that aligned with learners' interests. These adaptive capabilities are defining characteristics of why LLMs are so useful in contemporary e-learning settings.

Equally important is providing personalized experiences to all users regardless of their background. As illustrated in Figure 10, the system achieved fairness in distributed personalized feedback across different learner demographics. Measures of feedback equity and satisfaction remained consistently high among learners from the Global South, older adults, and non-English speaking participants—groups traditionally overlooked by automation systems.

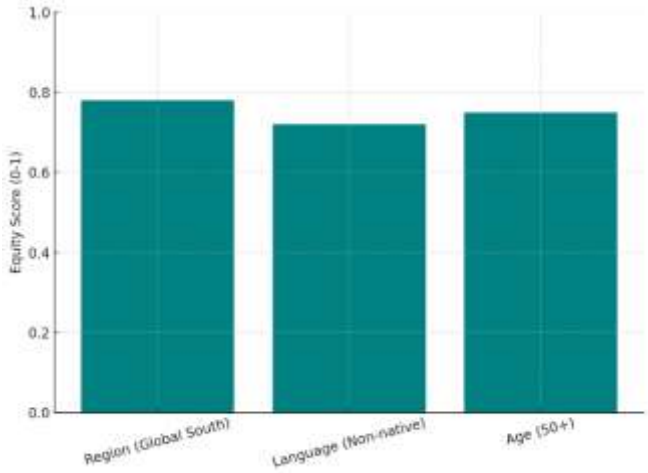

Fig. 10 Feedback Equity Across Learner Demographics (Region, Language, Age)

These outcomes corroborate not only the effectiveness of personalization but also the model's ability to promote inclusiveness and accessibility. When responsibly managed, LLMs have the potential to provide equitable feedback by ensuring that learners, irrespective of their geography, language, or age, receive high-quality, tailored responses crafted for their individual learning pathways.

5.3 Challenges in Integration with MOOC Platforms

Regardless of the encouraging outcomes, the incorporation of LLMs into current MOOC frameworks presents several concerns. One major issue stands out—computational cost. As discussed previously, large-scale GPT-4 implementation faces particularly harsh institutional funding constraints due to the token-based pricing paradigm. Although GPT-3.5 offers a more affordable option, it also demands substantial cloud resources, especially in real-time or high-concurrency environments.

Another challenge in implementation is the compatibility of content. Not every type of assignment lends itself to AI evaluation. For example, tasks that require highly subjective self-contextualization, effective storytelling, and ethics deliberation in given context discussive frameworks may be too advanced even for highly sophisticated LLMs. Along with this, as much as models can cope with multilingualism, the sinking incisiveness of fluency in lower-resourced languages hampers freely global deployment.

Infrastructure challenges also appear with simultaneous active users. Feedback efficiency gain in relation to the concurrent users is shown in Figure 11. The curve indicates that efficiency gain initially grows proportionally to scale. However, a point is eventually reached when infrastructural constraints, throttling of APIs, and sudden increases in latency would start lowering returns.

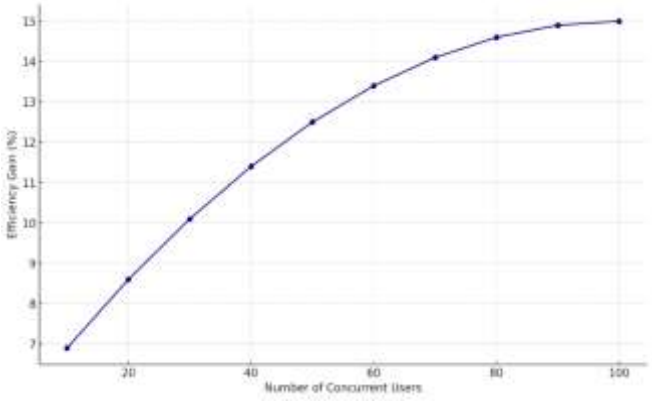

Fig. 11 Feedback Efficiency Gain vs Number of Concurrent Users

As shown, this pattern demonstrates the deliberate architectural need sought after. Institutions need to accommodate not only the average load but also the peak usage during exam periods and deadlines for assignments. Risk mitigation strategies like feedback queueing, hybrid cloud implementation, and model caching can address the problem but require additional resources for refinement and sustained help.

A last issue addresses the governance of cross platform integration concerning hierarchical feedback. Because feedback is sensitive to educational guidelines, institutions must create clear review, audit, and edit protocols pertaining to content LLMs produce. In the absence of an instructor- AI feedback loop, there is a potential danger of pedagogical drift—erosion of instructional quality over time.

5.4 Socio-technical Considerations and Bias Risks

Integrating LLMs in education brings forth socio-technical risks and ethical considerations that need special attention. LLMs are created using enormous datasets that are generally not curated, containing structural biases, inequities, and prevailing societal norms. Without implementation of active bias mitigation techniques, these datasets reproduce or reinforce biases.

Some LLM-assigned roles or conversational contexts were identified during the experiment in which learners' cultural or intention-related contextualized scaffolds were incorrectly identified. Some pieces of feedback subtly endorsed western-centric academic norms alienating learners from non-western indigenous traditions regarding arguments and expression. While these concerns were infrequent, they reveal a more troubling aspect of using sophisticated, general-purpose models in deeply-contextualized educational environments.

In addition to content bias, the risk of AI reliance also exists. With effective feedback systems in place, there is a danger of instructors avoiding the feedback loop which stops engagement with content. This gives birth to a new risk as feedback may become repetitive, monotonous, or out of alignment with the course changes without human oversight. In addition, students may presume authority surrounding AI-augmented feedback especially when it is polished, refined, or framed in formal syntax.

Such systems raise additional privacy and data protection issues. Anonymized datasets governed with strict policy frameworks will be required. Access to learner submissions at scale may allow for supporting or personally identifiable information to be inadvertently disclosed. Sensitive information may also be disclosed. In compliance with legal frameworks such as GDPR, FERPA, and privacy laws, clearer data governance, anonymization protocols, and opt-out options must be established.

Finally, the use of LLMs L raises ethical considerations surrounding authorship, instructional agency, and the presence of a human educator. The divide between teaching and machine instruction begins to soften when machines administer emotional affirmation and critique. We should always try to control and be careful of technologies that substitute some elements of teaching that ought to be retained—the relational, empathetic, critical, and deeply human dimensions of teaching. These technologies should ideally be able to supplement, not replace, educators' authentic interactions with students.

VI. CONCLUSION

6.1 Summary of Contributions

This research examined the capabilities of LLMs, particularly GPT-3.5 and GPT-4, in augmenting feedback mechanisms within MOOCs. With systematic scrutiny spanning several disciplines and thousands of learner engagements, our studies showed that LLMs achieved remarkable feedback construction that was insightful and contextually relevant in relation to instructors' expectations.

In particular, GPT-4's feedback was rated highly with respect to BLEU and ROUGE scores as well as overall feedback from peers and instructors. In addition, GPT-based systems drastically outperformed the traditional NLP engines in speed, usefulness, and breadth of assessment. AI feedback resulted in learners becoming more motivated and satisfied with the tasks at hand, enhancing active participation with assignments.

Improvements extended beyond the technical aspects too. The models demonstrated sophisticated personalization in tailoring responses considering the content, tone, and learner level which is impossible with rule-based feedback systems. In addition, the system seems to have the capability to provide equal coverage in feedback to members from different demographic groups, thus suggesting potential for deployment at scale for increased inclusivity.

6.2 Practical Implications for MOOC Platforms

These results directly affect the strategies of MOOC providers specifically in the areas of engagement and learner support. Conventional feedback systems based on instructor

or peer review are too slow, inconsistent, and unscalable. These are some of the gaps that LLMs can fill by providing feedback that is automated, timely, and responsive to individual learner requirements.

For educators, this translates to less time spent on grading and more time spent mentoring learners or developing instructional materials. For the platforms, it translates to better learner retention and higher rates of course completion. LLMs can be incorporated into existing LMS frameworks via APIs, providing responsive feedback in real time across different content, subjects, and formats, professionally and timely.

Cost is still a concern. The best performance comes from GPT-4, but its operational cost is steep. GPT-3.5 is a stronger middle option, balancing quality and efficiency. A blended approach can be adopted where Institutions save GPT-4 for more complex assignments requiring deeper insight and use lighter models for more straightforward tasks.

Privacy and data protection policies must receive equal attention. Platforms must comply with data protection policies concerning learner identity feedback systems require submission due to giving structured feedback identity protection, anonymization, and audit trails policy regulation as well as implement obscured processes where identities are hidden and auditability policies ensuring that records can be checked are put in place.

6.3 Recommendations for Adoption

To mitigate social, ethical trust, and safety risks, the following best practices are suggested for integrating LLMs into the MOOC framework. First, implement gradual steps. For onboarding, use AI-provided feedback on formative assessments. This will help ease both learners and instructors into the system and build confidence. Expand scope to more critical tasks later.

Second, delay full automation. Even the most modern of models require instructor approval for fault-sensitive issues or evaluations, as some topics need critical gaze attention. With instructor-in-the-loop models, quality control is preserved and system behaviour can be adjusted over time.

Third, provide justification. Reporting should explain why learners did not receive personalized feedback, especially when an AI system is used. Learners must have the ability to contest erroneous or inconsistent feedback. Comprehensible systems offer accountability while autonomous systems uplift learners.

Lastly, inclusivity must be deprioritized. Prioritize testing during refinement to keep bias out of the fine-tuning process. Test across demographics with divergent models. All levels of AI as an educational tool must be just, transparent, and equitable.

To conclude, usage of LLMs opens up many opportunities on automating feedback features in MOOCs. These models can be leveraged to greatly improve learner experiences, enhance instructor support, and outcomes at scale educational outcomes when accompanied with appropriate ethical frameworks and technical infrastructure.

VII. FUTURE WORK

This research did demonstrate a promising use of Large Language Models towards automating feedback on MOOCs, however, multiple aspects still require attention in both research and development. One such aspect includes LLMs's feedback diversity and cultural adaptability. Subsequent versions of the LLMs should be trained on data encompassing a broader spectrum of subjects, languages, and learner demographics to make the feedback more relevant, contextually sensitive, and pedagogically appropriate. Also, LLMs's multilingual feedback generation capabilities need attention, particularly for learners situated in non-English speaking countries where performance is lacking. Furthermore, providing LLMs with the ability to respond more human-like by, for example, emotionally recognizing learners' encouragement or frustration could enhance make the responses appear more supportive.

The use of LLMs in chat support services or AI assistants with MOOCs classes are other areas with great potential. Integrating feedback systems together with conversational agents would allow learners to actively engage in error correction and error reflection on a deeper level. Furthermore, exploring how LLMs can be incorporated into learning systems that not only provide feedback, but adapt the instruction content based on the performance of the learner, would be beneficial in future LLM research. For safe and sustainable use, however, addressing issues such as scaling up the infrastructure, lowering token costs, and using bias mitigation through reinforcement learning will be essential. Overall, with the advancement of LLMs, the scope of their application in intelligent tutoring systems and digital education will increase, which means that both educators and decision makers will need to work together with technology experts constantly.

REFERENCES

[1] Algahtani, A. M. I. R. A. H. (2024). A Comparative Study of AI-Based Educational Tools: Evaluating User Interface Experience and Educational Impact. *Journal of Theoretical and Applied Information Technology*, *102*(5).

[2] Alraimi, K. M., Zo, H., & Ciganek, A. P. (2015). Understanding the MOOCs continuance: The role of openness and reputation. *Computers & Education*, *80*, 28-38.

[3] Anandhi, S., Rajendrakumar, R., Padmapriya, T., Manikanthan, S. V., Jebanazer, J. J., & Rajasekhar, J. (2024). Implementation of VLSI Systems Incorporating Advanced Cryptography Model for FPGA-IoT Application. *Journal of VLSI Circuits and Systems, 6*(2), 107–114. https://doi.org/10.31838/jvcs/06.02.12

[4] Basri, W. S. (2024). Effectiveness of AI-powered Tutoring Systems in Enhancing Learning Outcomes. *Eurasian Journal of Educational Research (EJER)*, (110).

[5] Blasco, A., & Charisi, V. (2024). The Impact of Large Language Models on Students: A Randomised Study of Socratic vs. Non-Socratic AI and the Role of Step-by-Step Reasoning. *Non-Socratic AI and the Role of Step-by-Step Reasoning (December 02, 2024)*.

[6] Castiñeira, M., & Francis, K. (2025). Model-driven design approaches for embedded systems development: A case study. *SCCTS Journal of Embedded Systems Design and Applications, 2*(2), 30–38.

[7] Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023, September). The impact of ChatGPT on higher education. In *Frontiers in Education* (Vol. 8, p. 1206936). Frontiers Media SA.

[8] Fauvel, S., Yu, H., Miao, C., Cui, L., Song, H., Zhang, L., ... & Leung, C. (2018). Artificial Intelligence Powered MOOCs: A Brief Survey [Paper presentation]. In *IEEE International Conference on Agents. Singapore*.

[9] Ibrahim, R. (2020). Workstation Cluster's Hadoop Distributed File System Simulation and Modeling. *International Journal of Communication and Computer Technologies, 8*(1), 1-4.

[10] Jia, Q., Cui, J., Du, H., Rashid, P., Xi, R., Li, R., & Gehringer, E. (2024). LLM-generated Feedback in Real Classes and Beyond: Perspectives from Students and Instructors. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 862-867).

[11] Jiang, Z., Lin, H., Zhong, Y., Huang, Q., Chen, Y., Zhang, Z., ... & Liu, X. (2024). {MegaScale}: Scaling large language model training to more than 10,000 {GPUs}. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)* (pp. 745-760).

[12] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[13] Khalil, H., & Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-A literature review. *EdMedia+ innovate learning*, 1305-1313.

[14] Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, *355*(6322), 251-252.

[15] Liu, F., Wang, Y., Feng, Q., Zhu, L., & Li, G. (2024, August). Optimizing E-Learning Environments: Leveraging Large Language Models for Personalized Education Pathways. In *2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024)* (pp. 811-817). Atlantis Press.

[16] Moore, R. L., & Blackmon, S. J. (2022). From the learner's perspective: A systematic review of MOOC learner experiences (2008–2021). *Computers & Education*, *190*, 104596. https://doi.org/10.1016/j.compedu.2022.104596

[17] Omar, M., Choi, S., Nyang, D., & Mohaisen, D. (2022). Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, *10*, 86038-86056.

[18] Prasath, C. A. (2023). The role of mobility models in MANET routing protocols efficiency. *National Journal of RF Engineering and Wireless Communication, 1*(1), 39-48. https://doi.org/10.31838/RFMW/01.01.05

[19] Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied sciences*, *13*(9), 5783. https://doi.org/10.3390/app13095783

[20] Silva, J. C. da, Souza, M. L. de O., & Almeida, A. de. (2025). Comparative analysis of programming models for reconfigurable hardware systems. *SCCTS Transactions on Reconfigurable Computing, 2*(1), 10–15.

[21] Stamper, J., Xiao, R., & Hou, X. (2024, July). Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education* (pp. 32-43). Cham: Springer Nature Switzerland.

[22] Wei, L., & Lau, W. C. (2023). Modelling the power of RFID antennas by enabling connectivity beyond limits. *National Journal of Antennas and Propagation, 5*(2), 43–48.

[23] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 1-27.

[24] Zhu, M., Sari, A. R., & Lee, M. M. (2020). A comprehensive systematic review of MOOC research: Research techniques, topics, and trends from 2009 to 2019. *Educational Technology Research and Development*, *68*, 1685-1710. https://doi.org/10.1007/s11423-020-09798-x