Grenze International Journal of Engineering and Technology, June Issue

Performance and Scalability Analysis of AI-Driven Cyber-Secure Assessment Platforms

Saksham Pokhrel[1], Saurab Gurung[2], Renuka Chaudhary[3] and Sushant Jhingran[4]
[1-4]Department of Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India
Email: sakshampokhrel175@gmail.com, saurb21112002@gmail.com, renu16516@gmail.com, sushantjhingran@gmail.com

***Abstract*— The rapid growth of digital learning has put a strain on scalable and secure online exam systems that will be able to perform in a steady fashion when there are large numbers of concurrent users. Proctoring using Artificial Intelligence (AI) can enhance academic integrity by providing continuous monitoring, behavioral analysis and real-time surveillance, but, because AI inferences are persistent, the system is likely to be limited in its scalability and reliability. This paper presents a performance and scalability analysis of an AI-adapted cyber-secure online examination system, which is developed based on a modular full-stack design, stateless backend services, asynchronous processing, and container deployments. Standard load-testing and benchmarking methods were used to carry out experimental analysis at the different workloads. The core performance mesures that were evaluated including reaction time, processing capacity, resource use and Ai inference delay. Result shows that the proposed system will have low latency and high throughput with constant AI-based monitoring at low to moderate concurrency and show the predictable performance degradation at the high loads. The result indicates the balance of the inconsistent AI monitoting and expandability and also offers actionable information in the development of efficient, safe and scalable web based assessment platforms.**

***Index terms*— AI-based Online Examination, Cyber-Secure Assessment, Performance and Scalability, Automated Proctoring, Resource Optimization.**

I. Introduction

With the rapid growth of e-learning, distance learning and computer-based teaching methods, Online Evaluation systems have assumed a vital role in modern academic appraisal which enables the process of evaluation to take place at any place and any time. Behaviour Identification, Activity Tracking, Face Recognition and video tracking in real-time function as an AI-based method that is typically applied to promote academic integrity through preventing impersonation and unauthorized behaviour [1], [8]. Online assesments are the type of cyber systems which are sensitive to security because these AIs features enhance the security of the examinations with means of constant monitoring and automatic response to suspicious behavior. However, continous AI observation imposes significant computational loads on the systems which affects the systems in terms of usability, scalability and resource efficiency [2], [9]. Current Assessment frameworks based on the internet are often unable to find a balance between an effective security approach and a reliable performance when a large concurrency is involved [6], [7].

Grenze ID: 01.GIJET.12.2.300_20
© Grenze Scientific Society, 2026

The lightweight systems provide even better performance but with fewer monitoring and the system that have high monitoring face deterioration in performance as the number of user's loads increase [2], [3]. Although the effectiveness of AI-based cheating detection has been sufficiently studied, sacability and performance of AI tasks on large scale environments have not been thoroughly studied [9], [14].

II. Literature Review

The growing use of online examination systems have resulted in the research into the enhancement of acessibility and academics honesty. The earlier online systems have only simple web based assessment systems like timer features, randomization of questions and automatic scoring features but has no efficiency surveillance thus was prone to impersonation and deciet [1], [4], [5]. To solve these problems, AI-powered procotoing features like face recognition, eye tracking, head tracking and tab switching are proposed to improve the level of exam security. The majority of the available research applies to these systems on the basis of the detection rates of cheating and the false positives [2], [14], and some research is devoted to the secure system design in terms of authentication, role-based access control, and secure communication systems [10], [11]. Although these developments have taken place, studies of the performance and scalability of AI-driven examination systems are still scarce. The high computational demands of continual AI applications, like real-time video surveillance, are relative to the usual event-driven system [2], [3]. Even though the concept of scalability such as load balancing and horizontal scaling has been extensively researched in the e-learning platforms, the extra effect of continuous AI inference is likely to be ignored [6], [9]. In addition, most studies are based on controlled settings that are not representative of large-scale examination conditions of the real world [1], [2].

III. System Architecture

The architecture design comprises the use of Angular-based single-page frontend, Spring Boot Backend, component-based AI proctoring, and containerized database to ensure scalability and consistent performance. Frontend deals with exam interfaces, timer, tracking user activity and the basic proctoring controls and improvements such as lazy loading and good state managements keep performance steady and without exception as many users as possible are using it. The backend is stateless which enables the resources to be horizontally scaled by providing RESTful authentication, exam administration, response submission and event monitoring solutions using JWT authentication. The proposed AI-based proctoring system uses a hybrid and modular approach that allows achieving the balance between accuracy in monitoring and computational performance, allowing the system to scale independently and avoid conflicts of resources. The system uses a normalized MySQL database that is optimized by indexing and query optimization to perform efficient data retrieval and Docker-based containerization and orchestration to provide uniform deployment, resource isolation, dynamic scaling, and enhanced reliability to large-scale examination environments.

IV. Methodology

In this section, the experimentation conducted to evaluate the effectiveness, scalability and reliability of the proposed AI-based online assessment system is described. The research adopts a quantitative design because it will evaluate the performance of the system as the number of parallel users advances with the system performance constantly maintained due to the presence of AI-driven monitoring. The reason why AI-proctored environments need always-available computational resources is because it continuously detects the video and tracks the activities, as such, the assessment was conducted based on the principles of repeatability, controlled experiments, and quantitative benchmarking procedures as per the standards of distributed systems [15], [16]. Regarding performance evaluation, the system was split into the critical components, which constitute stateless back-end services that are used to carry out the authentication process, exam management, logging, and interaction with the database, which enables the system to scale horizontally in a scenario where there are many parallel requests, and an AI tracking component, which runs face recognition and behavioral analysis continuously, resulting in a continuous consumption of CPU and memory and directly impacting scalability [2], [9]. The scalability testing was conducted based on incremental load-testing process whereby, the number of simultaneous users was gradually increased and data was captured when steady-state load operations were maintained at each level of incremental load. It was examined on the basis of the evolution of response time, AI effect on throughput, the tendencies of CPU and memory consumption, and determining the saturation values, and performance limits that would have granted an estimation of the degradation curves and scalability limits [16], [18]. The performance data were collected by using application logs, monitoring tools, and resource tracking systems of the client, backend, and AI components and statistical measures of the mean, peak, and the variance were calculated in every load scenario [15], [17].

V. Performance and Scalability Evaluation

In this subsection, the runtime performance characteristics are assessed by testing the functionality of the real-time system when it is in active operation and the interaction between the frontend, backend, and AI monitoring parts is evaluated. The persistence in AI application would put long-term computational load on the system that would adversely affect its efficiency without proper management [1], [2], and online tests would demand a stable and predictable execution performance [8]. Responsiveness refers to the system's response to user interaction, including page loading, navigation and submission of answers under the conditions of concurrent user loads where the delay patterns are observed to be within the acceptable performance limits because AI-based monitoring can add latency to peak loads as the same resources are used [9], [14]. The experiments were deployed in web-based setting within a simulated environment of real institutional deployment employing a client-server structure, stateless backend services, RESTful APIs and ongoing AI monitoring. Simulated concurrent users were used to test scalability and measured metrics of the performance such as response time, AI inference latency, CPU usage, memory usage, throughput and error rate were measured using monitoring tools. Experiments were repeated in steady-state conditions to provide proper performance and scalability evaluation.

VI. Result and Performance Analysis

A. Response Time Analysis

When the number of users was 200, the average system response time was approximately 178 ms. The latency also marginally went to 185 ms when 400 users were serving which means that the backend is offering decent request management and scalability. Response time increased sharply at 500 users, to approximately 530 ms, which is due to the resource being saturated due to CPU contention between running AI inference. The breakdown was slow but not sudden, and it was a challenge to the stability of the system behavior at the time of stressing it.

B. Reliability and Error Analysis

All experiments had a low error rate (= 0.01% at most), and the system did not experience any failures, errors during requests, or instability. It also validates the strength of the architecture in the event that it is exposed to intense computational workload.

C. Throughput Analysis

The throughput rose according to the low to moderate loads and peaked at the 228 requests/sec with 200 users. The throughput was constant at 218 requests/sec, with 400 users, which is the evidence of the efficient scaling. It decreased to 187 requests/sec with 500 users which is a resource limitation issue in the back-end, though the received requests were handled in a dependable manner without any failure.

D. Resource Utilization Analysis

The CPU consumption was rising gradually with the user load reaching about 90 percent with 500 users at a time, with the CPU resources seen to be the scaling limit. The use of memory was stable that showed that there was no memory leakage and there was proper management of memory. These findings validate the fact that ongoing AI inference generates unremitting computational demands that directly affect scalability demands.

E. Summary of Performance Results

Overall, the AI-based evaluation system showed an overall consistent performance with low to moderate concurrency. Continuous AI monitoring added additional load with higher loads, but the system remained responsive and reliable to high levels of concurrency with limited visibility on scalability constraints.

TABLE I. PERFORMANCE METRICS VS. USER LOAD

Concurrent Users	Avg Response Time (ms)	Throughput (req/sec)	Error Rate (%)	CPU Utilization (%)	Performance Interpretation
200	178	228	0.01	58	Excellent
400	185	218	0.00	82	Very Good
500	530	187	0.00	90	Saturation Begins

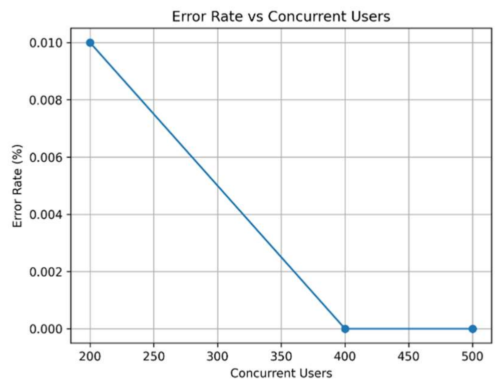

Fig. 1. Error rate vs. concurrent users showing improved system stability under higher load

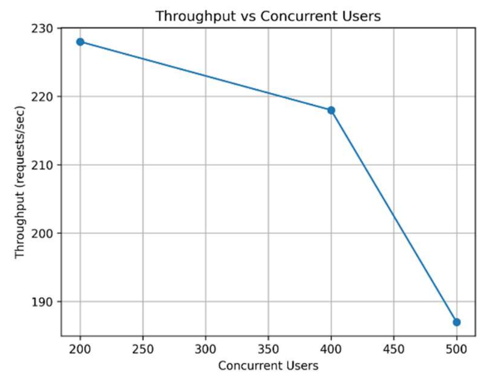

Fig. 2. Throughput vs. concurrent users indicating performance reduction at heavy load

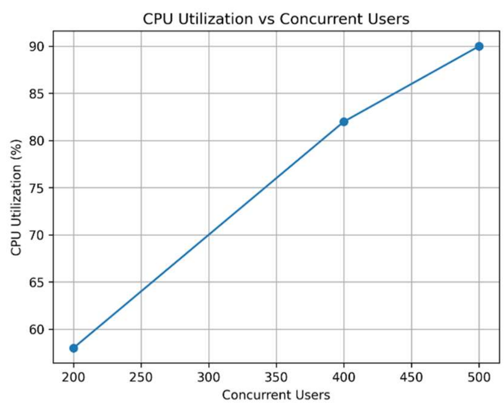

Fig. 3. CPU utilization vs. concurrent users showing higher resource usage under heavier load

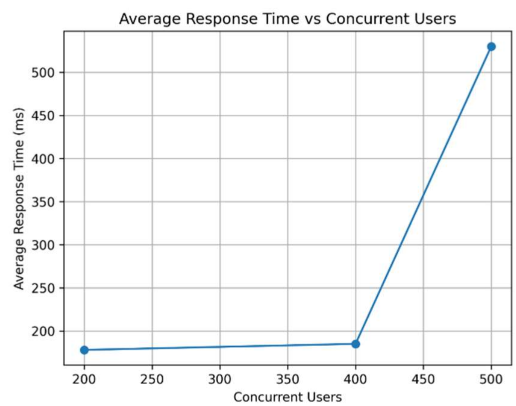

Fig. 4. Average response time vs. concurrent users illustrating performance degradation at high concurrency

VII. DISCUSSION

The research results of the experiment provide clear insight into the effectiveness and scalability of the AI-based online evaluation system in the context of continuous monitoring. As shown in Table I and Figs. 1–4, there is a trade-off between maintaining academic integrity through AI-based proctoring and the computational cost of real-time AI processing. Analysis of response time shows that there is a steady and efficient response time at light to moderate loads, and that latency increases gradually with 200 to 400 users, indicating that it is a well-behaved backend with high scalability and utilization of resources. However, response time grows exponentially with higher concurrency (500 users) because of increased competition over common computational resources because of asynchronous inference of AI. Throughput analysis (Fig. 2) shows that there is good request management with moderate levels, but critical declines below high concurrency, as a result, the computational cost of AI oversight is increasing. The trends in CPU utilization (Fig. 3) support the fact that processing resources are the primary scalability constraint, as CPU utilization approaches a point of saturation and negatively affects response time and throughput. On the other hand, steady usage of memory is an indication of a good memory management meaning that any reduction in performance is first and foremost constrained by computation. Despite the workload increment, the low error rates (Fig. 1) suggest that there is a high system reliability and operational stability.

Overall, the system can be run at average workloads consistently, but there are scalability problems due to the processing demands of continual AI inference. These findings suggest that AI needs improvement, the utilization of resources should be optimized and architectural solutions such as hardware acceleration or distributed computing should be embraced to implement AI in large scales.

VIII. Conclusion and Future Work

The study presented a comprehensive evaluation of the functionality and scalability of an AI-based cyber-safe online review platform that has been used and under constant-real time monitoring. There were experimental tests that pointed to the fact that the proposed architecture is capable of supporting high responsiveness, predictable throughput, and reliable performance at the low to moderate loads by the users. Stateless backend infrastructure and modular AI monitoring system along with better design enable to control the requests and maintain the integrity of exams through continuous proctoring using the assistance of AI. These findings can prove that the system can balance the requirements of security and efficiency of performance in the moderate working conditions. Nonetheless, higher levels of scalability to greater concurrency are also demonstrated to be limited by the assessment. The more users there are at once, the slower the response time and the slower the throughput because of the computational burden of inference by continuous AI. A large CPU consumption indicates that the processing capacity is the most limiting factor of scalability and the steady memory consumption indicates that the resources were managed successfully and that the system is reliable. It has been observed that even with the increment in the requirement of the computers, there is still low error rate which indicates that the architectural reliability is high and it is a trade-off between rigorous AI control and scalability.
Future directions will involve improving scalability and computational efficiency by making AI inference as cheap as possible by simplifying model optimization, running with GPUs, and edge or distributed computing devices. Other enhancements, like cloud-native auto-scaling, resource allocation and realistic performance and scale testing using network-latency and a broad selection of client environments can be made to make the system even more performant and scalable.

References

[1] A. Tweissi, W. Al Etaiwi, and D. Al Eisawi, “The Accuracy of AI-Based Automatic Proctoring in Online Exams,” Electron. J. e-Learn., vol. 20, no. 4, 2022, doi: 10.34190/ejel.20.4.2600.

[2] T. Singh, R. R. Nair, T. Babu, and P. Duraisamy, “Enhancing Academic Integrity in Online Assessments: Introducing an Effective Online Exam Proctoring Model Using YOLO,” Procedia Computer Science, vol. 235, pp. 1399–1408, 2024, doi: 10.1016/j.procs.2024.04.131.

[3] N. Y. Al Hoqani, T. Regula, S. G. K. Kumar, M. R. Battina, A. N. Al Salmi, and K. Ayyaraju, "Artificial Intelligence-Enabled Online Exam Proctoring Systems: Technologies, Challenges, and Scalable Solutions for Remote Education," 2025 IEEE International Conference on Computation, Big-Data and Engineering (ICCBE), Penang, Malaysia, 2025, pp. 179–183, doi: 10.1109/ICCBE65177.2025.11255752.

[4] R. Raja Subramanian, B. Surekha, P. SubbaRao, G. A. Kesava Chowdary, and M. Venkatesh, "EduGuard – AI Based Online Exam Management System," 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 2025, pp. 1-6, doi: 10.1109/ICCRTEE64519.2025.11052922.

[5] P. A, P. M. N, and S. A. Alex, "Deep Learning Classroom Management System for Exam Monitoring and Student Behavior Detection," 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 2024, pp. 2004-2009, doi: 10.1109/I-SMAC61858.2024.10714864.

[6] S. Kadam, M. Kothalkar, and D. Ambawade, "Blockchain-Enabled Examination Platform: A Secure Approach for Academic Assessments," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725734.

[7] P. H. Sharma and P. Singh, "CDCEF: A Cloud-Based Data Volatility & CSP Reliance Eradication Framework," 2024 17th International Conference on Security of Information and Networks (SIN), Sydney, Australia, 2024, pp. 1-6, doi: 10.1109/SIN63213.2024.10871550.

[8] M. S. Hamadi et al., "Mapping the Characteristics of Web Services to Queueing Theory Models for Performance Evaluation," 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2025, pp. 1-8, doi: 10.1109/ICBATS66542.2025.11258372.

[9] N. N. Nayim, A. Karmakar, M. R. Ahmed, M. Saifuddin, and M. H. Kabir, "Performance Evaluation of Monolithic and Microservice Architecture for an E-commerce Startup," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023, pp. 1-5, doi: 10.1109/ICCIT60459.2023.10441241.

[10] S. K. Kotakonda, "Benchmarking Container Orchestration Platforms: A Comparative Analysis of Kubernetes and AWS ECS for Stateful Microservices," 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2026, pp. 1-5, doi: 10.1109/ICAIC67076.2026.11395772.

[11] A. Malhotra, A. Elsayed, R. Torres, and S. Venkatraman, "Evaluate Canary Deployment Techniques Using Kubernetes, Istio, and Liquibase for Cloud Native Enterprise Applications to Achieve Zero Downtime for Continuous Deployments," in IEEE Access, vol. 12, pp. 87883-87899, 2024, doi: 10.1109/ACCESS.2024.3416087.

[12] V. Tiwari, S. Upadhyay, J. K. Goswami, and S. Agrawal, "Analytical Evaluation of Web Performance Testing Tools: Apache JMeter and SoapUI," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 519-523, doi: 10.1109/CSNT57126.2023.10134699.

[13] M. Hunko, A. Kovalenko, and V. Tkachov, "Comparative Analysis of Cloud and Fog Computing: Performance and Suitability in Different Scenarios," 2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2025, pp. 1-6, doi: 10.1109/KhPIWeek61436.2025.11288655.

[14] B. Naets, W. Raes, R. Devillé, C. Middag, N. Stevens, and B. Minnaert, "Artificial Intelligence for Smart Cities: Comparing Latency in Edge and Cloud Computing," 2022 IEEE European Technology and Engineering Management Summit (E-TEMS), Bilbao, Spain, 2022, pp. 55-59, doi: 10.1109/E-TEMS53558.2022.9944509.

[15] V. Melikyan et al., "Cross-Platform Evaluation of Visual Localization Models: Benchmarking CPU, GPU, Jetson, and FPGA Implementations," 2025 IEEE East-West Design & Test Symposium (EWDTS), Tbilisi, Georgia, 2025, pp. 1-4, doi: 10.1109/EWDTS67441.2025.11303675.

[16] J. Sun, X. Yi, A. Symons, G. Gielen, L. Eeckhout, and M. Verhelst, "An Analytical Model for Performance-Carbon Co-Optimization of Edge AI Accelerators," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 45, no. 3, pp. 1499-1503, March 2026, doi: 10.1109/TCAD.2025.3602746.