Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 2 (2026) pp: 4407-4415
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Development of Automated Essay Scoring Using Retrieval Augmented Generation in SAGE

I Gusti Nyoman Sapta Wiguna, Ida Bagus Nyoman Pascima, Luh Putu Eka Damayanthi
Program Studi Pendidikan Teknik Informatika, Fakultas Teknik dan Kejuruan, Universitas Pendidikan Ganesha
sapta@student.undiksha.ac.id*, gus.pascima@undiksha.ac.id, ekadamayanthi@undiksha.ac.id

Abstract

Academic assessment through essay questions is a fundamental component of the educational ecosystem designed to measure students' cognitive depth and critical reasoning. However, manual essay grading presents significant pedagogical and administrative challenges including high susceptibility to subjective bias and an overwhelming workload for educators. To address these critical issues this research develops the Smart Automated Grading Engine or SAGE, an advanced Learning Management System engineered to automate qualitative assessments. SAGE integrates large language models via the Gemini API with a robust Retrieval Augmented Generation architecture. By strictly grounding the artificial intelligence evaluation process in teacher curated reference documents and specific grading rubrics the system effectively neutralizes the risk of information hallucination. The system was empirically validated at SMAN 1 Blahbatuh involving 180 authentic essay responses from 36 eleventh grade students. The automated assessments were statistically compared against the manual evaluations of three expert history teachers. Comprehensive technical evaluations utilizing Black Box and White Box testing confirmed the platform absolute functional stability and architectural security. Crucially the accuracy testing demonstrated exceptional pedagogical reliability where the SAGE platform achieved a Quadratic Weighted Kappa coefficient of 0.9133 categorizing its performance as having almost perfect agreement. Furthermore, the system exhibited a remarkable precision rate of 94.44 percent within a stringent 10 point score tolerance. Ultimately the integration of this technology proves to be an effective objective and efficient solution capable of replicating human evaluation sharpness while significantly alleviating educator burnout.

Keywords: *Automated Scoring, Learning Management System, Retrieval Augmented Generation, Gemini API, Assessment Accuracy*

1. Introduction

Academic assessment is a central pillar in the educational ecosystem that serves to measure the extent to which student competencies and comprehension have been achieved. Among the various evaluation instruments available, essay questions are considered the most effective method because they can comprehensively dissect cognitive depth, critical reasoning abilities, and student written communication skills. Unlike multiple choice questions, essays require students to construct their own thoughts in a logical and structured manner [1].

However, in practice, manual essay assessment still leaves major challenges that often hinder educational objectivity. The main problems that frequently arise are high levels of subjectivity and potential inconsistency among raters. Furthermore, the extremely heavy workload of educators coupled with narrow grading time limits often result in a decline in the quality of feedback provided to students [2]. This human fatigue factor becomes a loophole that can undermine the fairness of academic assessment in schools.

To circumvent the pervasive challenges of manual evaluation, researchers have extensively explored the paradigm shift toward automated assessment solutions powered by artificial intelligence. Recent systematic cross disciplinary surveys underscore that leveraging powerful large language models for grading can dramatically reduce the time educators spend on evaluation tasks. By automating the most time consuming aspects of grading, educators can reallocate their focus toward more interactive teaching activities and personalized student mentoring. However, the implementation of such advanced technology is not merely about algorithmic automation. Recent evidence strongly indicates that integrating artificial intelligence within a structured pedagogically centered process, rather than attempting to fully replace human educators, is what ultimately yields a reliable, valid, and

sustainable approach to automated essay scoring in dynamic educational settings [3]. This human in the loop perspective ensures that automated systems act as robust pedagogical assistants.

The integration of technology through supporting platforms or Learning Management Systems becomes highly crucial as a solution to these problems. The implementation of learning management systems has been proven to be capable of facilitating educational evaluation needs and significantly improving the efficiency of the teaching and learning process[4]. A properly structured system allows educators to manage tasks and assessments in a more organized manner.

Along with the rapid advancement of artificial intelligence technology, the use of large language models is beginning to be implemented specifically for automated scoring within the educational ecosystem. The integration of large language models with the Retrieval Augmented Generation method is highly effective in automating the evaluation process within learning platforms with high accuracy [5]. The use of adaptive natural language processing in educational information systems also plays an important role in eliminating subjectivity bias in a measurable way [6].

This RAG method bridges the technical constraints of standard language models that often experience information hallucination, ensuring that factual accuracy is maintained in educational applications [7]. This approach allows the system to first conduct a search on reference documents and rubrics that have been curated by educators before providing evaluation results [8]. This experiment focuses on the development and performance testing of the SAGE platform as a form of precise automated essay assessment implementation. Through the utilization of the Gemini API combined with this architecture, every generated score will have a valid data foundation that is relevant to the learning materials. The main focus of this research is to test the extent to which the SAGE system can replicate the sharpness of human teacher assessments consistently and objectively through an approach based on curated data at SMAN 1 Blahbatuh.

2. Research Methods

This research applies the Research and Development method to design and validate an efficient automated assessment system [9]. The system development workflow is entirely guided by the Agile model which includes requirements analysis, design, development, and testing phases to ensure each cycle produces an iteratively tested product. The subjects in this study involve 36 students and three expert history teachers at SMAN 1 Blahbatuh as the primary location for the experiment. The research object is focused on the Smart Automated Grading Engine or SAGE platform which integrates Retrieval Augmented Generation technology for automated essay scoring.

Figure 1. Agile Model Workflow

The research procedure begins with the requirements identification phase where researchers collected primary data through in-depth interviews with educators. This step aims to map the technical constraints in the manual grading process and establish the required system functional specifications such as learning modules and the provision of automated assessment rubrics. The results of this needs identification are then used as the main foundation in structuring the system framework to align with practical needs in the school environment. The next phase is design which involves creating the information system architecture and a structured user interface design. In this phase database schema design is conducted to guarantee data security and integrity. Researchers also mapped the artificial intelligence integration workflow through the utilization of the Gemini API combined with the Retrieval Augmented Generation mechanism so that the assessment process remains grounded in reference documents and

DOI: https://doi.org/10.31004/riggs.v5i2.8979

rubrics curated by educators. After the design is complete the research proceeds to the development phase to build the SAGE platform in a functional web-based application format. Development is carried out by applying programming logic that allows the system to automatically search for reference documents before the generative language model provides the final score. Technically the Retrieval Augmented Generation method works by extracting the most relevant information from uploaded material modules and teacher rubrics then inserting this context into the main processing instructions, a mechanism similarly utilized to extract and evaluate semantic similarities in project based learning environments [10]. Through this context insertion mechanism, the language model is strictly guided to formulate evaluations based only on the provided facts thus minimizing the risk of information hallucination. The main features implemented include an essay assignment collection module and the provision of comprehensive diagnostic feedback for students.

The procedure concludes with the system testing phase which is conducted comprehensively to measure functionality and accuracy. The testing encompasses Black Box functional testing to ensure all navigation interfaces and class management features run without operational failure, White Box testing to evaluate the internal program code structure and technical API integration stability, and accuracy testing using the Quadratic Weighted Kappa metric to measure the statistical consistency and agreement between the automated scores generated by the SAGE platform and the manual scores from the three history teachers on 180 student essay responses. The instrument grids used in this research are presented in Table 1, Table 2, and Table 3.

Table 1. Blackbox Testing Instrument

No	Feature Tested	Expected Output
1	Login	User is redirected to the main page based on account role.
2	Login	User remains on login page and sees an invalid credentials error.
3	Logout	Session ends and user returns to the login page.
4	Role Access	Access is denied (unauthorized/forbidden).
5	Role Access	Access is denied (unauthorized/forbidden).
6	Class - Join	Student successfully joins the class.
7	Class - Join	Join fails and class code not found message appears.
8	Class - Create	New class is created and appears in the class list.
9	Class - Edit	Class changes are saved and reflected in class details.
10	Class - Delete	Class is removed from the list and cannot be accessed.
11	Material - Create	Material is saved and appears in the material list.
12	Material - Edit	Updated material content appears on the student page.
13	Material - Delete	Material is removed from the section list.
14	Essay Question - Create	Question is saved and appears in the question list.
15	Essay Question - Validation	System rejects save and shows validation message.
16	Rubric - Create	Rubric is saved and can be used for grading.
17	Rubric - Validation	System rejects save and shows validation message.
18	Question Bank	Question appears in question bank with correct metadata.
19	Question Bank	Question is added to the class question list.
20	Submission	Submission is saved and status changes to submitted.
21	Submission	System rejects submit or marks it incomplete based on rules.
22	Submission Attempt	System rejects submit and shows attempt limit message.
23	Queue Grading	Status changes to queued and then processing.
24	Queue Grading	Status changes to completed and AI score is stored.
25	Queue Grading	Status changes to failed and error message is recorded.
26	Instant Grading	AI score appears immediately without queue wait.
27	Final Score Logic	Final score prioritizes teacher revised score.
28	Feedback Display	AI and teacher feedback are displayed correctly.
29	Score Report	Summary, distribution, and score table are displayed.
30	Score Report - Filter	Report data updates according to filters.
31	Score Report - Export	CSV file is downloaded and content matches displayed data.
32	Notifications	Notification is delivered to the relevant user account.
33	Calendar	Selected-date agenda displays correct events.
34	Data Security	Access is denied and no data is leaked.
35	Load Stability	System remains responsive and grading continues.

DOI: https://doi.org/10.31004/riggs.v5i2.8979

Table 2. Whitebox Testing Instrument

No	Evaluation Category	Expected Output
1	Architecture & Modularity	Module boundaries are clear and responsibilities are separated.
2	Architecture & Modularity	End-to-end flow is documented and traceable.
3	Architecture & Modularity	Components are modular and maintainable for future development.
4	Code Quality & Maintainability	Code is readable and consistent.
5	Code Quality & Maintainability	Validation and error handling are adequate.
6	Code Quality & Maintainability	Changes are localized and do not broadly impact unrelated modules.
7	Correctness (Scoring Logic)	Final score prioritizes teacher revised score over AI score.
8	Correctness (Scoring Logic)	Rubric calculations are consistent with rubric definitions.
9	Correctness (Scoring Logic)	System handles edge cases safely and predictably.
10	Data & Security	Access control is enforced correctly by role.
11	Data & Security	Queries are protected from potential SQL injection.
12	Data & Security	Sensitive data is not exposed to unauthorized roles.
13	Performance & Scalability	Endpoint remains efficient and responsive.
14	Performance & Scalability	System avoids excessive queries (no problematic N+1 pattern).
15	Performance & Scalability	Main request path remains responsive because grading is queued/asynchronous.
16	Testing & Quality Assurance	Critical scoring functions are covered by unit tests.
17	Testing & Quality Assurance	End-to-end API integration path is validated.
18	Testing & Quality Assurance	Coverage is adequate for the module's critical behavior.
19	Technical System Feasibility	System is feasible for real classroom-scale operation.
20	Technical System Feasibility	System is stable enough for regular essay assessment operations.

Table 3. QWK Score Interpretation

QWK Score Range	Interpretation	System Status
≥ 0.80	Almost Perfect Agreement	Acceptable
0.60 – 0.79	Substantial agreement	Acceptable
0.40 – 0.59	Moderate agreement	Needs improvement
0.20 – 0.39	Fair agreement	Not acceptable
< 0.20	Slight or no agreement (random like)	Rejected

The comprehensive data gathered through these testing instruments are subsequently analyzed to evaluate the overall performance and reliability of the SAGE platform. This structured evaluation ensures that the analysis of system functionality and scoring accuracy is conducted objectively to determine the success of integrating Retrieval Augmented Generation in the automated essay assessment process. The results obtained from these testing phases provide the factual foundation for the discussion on the technical effectiveness and pedagogical benefits of the system in a real world educational environment.

3. Results and Discussions

This section comprehensively details the outcomes of the Agile development phases applied to the Smart Automated Grading Engine platform and interprets the findings from the system evaluation. The implementation of the system is elaborated through the planning, design, development, and testing phases.

The initial phase in the Agile development of the Smart Automated Grading Engine platform is the requirements identification process which was conducted to accurately map the functional and technical specifications of the system. This step was executed through direct observation and in depth interviews with history educators at SMA Negeri Bali Mandara to understand the practical constraints in the academic evaluation process. The identification results revealed a significant phenomenon of extremely high educator workload in grading long essay answers manually which often led to a decrease in scoring consistency and unavoidable subjectivity. The necessity for a digital automated system in this educational environment aligns with the principle that implementing a structured digital learning approach within a Learning Management System is highly effective in facilitating educational needs and enhancing digital competence [11]. From these fundamental issues the primary functional requirements of the system were formulated including the capability to manage classes digitally upload learning materials as contextual references and perform automated assessments based on teacher rubrics. Educators require a system that not only provides raw numerical scores but also generates diagnostic feedback that is perfectly aligned with the ongoing history curriculum. This became the foundation for the developers to determine that the system must

DOI: https://doi.org/10.31004/riggs.v5i2.8979

dynamically process external documents before the artificial intelligence engine calculates the final score during the accuracy testing phase at SMAN 1 Blahbatuh. From a technical requirement perspective the system is obligated to possess an architecture capable of supporting the integration of a Large Language Model with a strict context control mechanism. The researchers established the requirement to use the Gemini API combined with the Retrieval Augmented Generation method to ensure the system does not solely rely on the general knowledge of the language model but specifically refers to the modules and rubrics uploaded by the educators. The integration of large language models with the Retrieval Augmented Generation method has proven to be highly effective in automating the evaluation process within learning platforms while significantly mitigating the risk of information hallucination [12]. Specifically by grounding the assessments in curated rubrics and instructor feedback [13]. This needs identification also covers aspects of data security and navigation simplicity for student users. All findings from this phase were subsequently documented as software specifications which serve as the primary guideline in the subsequent system design phase. The mature determination of requirements at the beginning of this Agile cycle ensures that the constructed platform truly becomes a relevant solution for pedagogical challenges.

The design phase of the SAGE system focuses on translating the requirement specifications into a comprehensive technical architecture. This process includes user interface design database modeling and mapping the artificial intelligence integration workflow. The interface design is conducted by prioritizing usability aspects so that students and teachers can operate the platform easily without requiring complex technical training. This approach is consistent with the principle that the development of interactive content must consider ease of navigation and the presentation of engaging information to support the effectiveness of digital learning [14]. In the technical aspect researchers structured a robust database schema using PostgreSQL to store primary academic data and Redis for session management to ensure responsive system performance when accessed concurrently. The database schema was designed through a normalization process to guarantee the security and integrity of student information and assessment rubrics curated by educators. The core of this design phase is the development of the large language model integration flowchart through the Gemini API combined with the Retrieval Augmented Generation engine.

Figure 2. System Flowchart

DOI: https://doi.org/10.31004/riggs.v5i2.8979

As illustrated in Figure 2 This architecture is designed in such a way that when an essay answer is submitted by a student the system will automatically perform a contextual search on the reference documents before formulating the final assessment instructions for artificial intelligence.

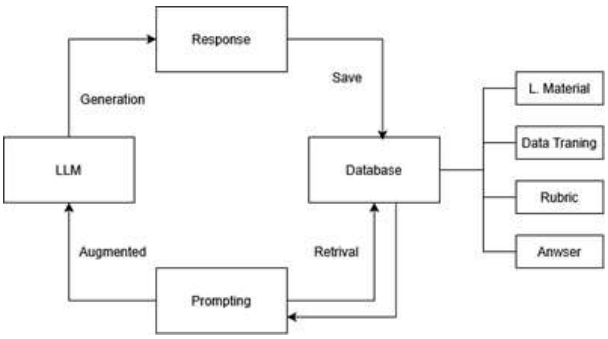

Figure 3. System Retrieveal Augmented Generation Ilustration

The design of the Retrieval Augmented Generation logic in SAGE, depicted in Figure 3, involves creating a database that stores representations of learning materials and teacher rubrics. This mechanism ensures that every evaluation generated has a strong factual basis and is highly relevant to the historical topics being tested. The implementation of this method in education has been proven to improve the response accuracy of language models while reducing the potential for information errors in the automated assessment process [12]. This entire architectural design becomes the blueprint in the development phase to ensure the SAGE system can operate stably and accurately when entering the testing phase at SMAN 1 Blahbatuh.

The development phase is the execution stage where the architectural blueprint and interface design are transformed into a fully functional software product. In this phase the Smart Automated Grading Engine platform was built using the Next js framework to produce an interactive and responsive web based user interface for both students and teachers. Meanwhile the server side processing logic was specifically developed using the Go programming language which offers exceptionally high performance in handling concurrent computational loads. For data management aspects the system implements PostgreSQL as the primary database to permanently store academic information along with Redis which functions to manage user sessions ensuring that access to the platform remains fast and stable. The utilization of these robust technologies aligns with the requirement that structured data modeling and reliable back end processing are crucial for developing an information system capable of handling complex academic workflows seamlessly [15]. The primary innovation focus during this development phase lies in coding the integration logic for the Retrieval Augmented Generation architecture. Developers constructed a complex algorithm connecting the internal system database with the Gemini API service. When history teachers upload assessment rubric documents and material modules the system processes all the text into contextually searchable data. Subsequently when a student clicks the submit button on the essay answer sheet the program flow automatically dissects the answer searches for the most relevant material intersections from the teacher reference documents and inserts this context into the main processing instructions before sending them to the language model. The technical implementation at this code layer successfully created a grading engine that strictly limits the operational scope of the artificial intelligence. The language model is forced to only extract scores and formulate diagnostic feedback purely based on the educator guidelines without fabricating any fictitious information. The effectiveness of integrating the Retrieval Augmented Generation architecture within a Learning Management System has been scientifically validated to automate student evaluations while maintaining factual integrity [5]. The profound advantage of deploying the Retrieval Augmented Generation architecture within the SAGE ecosystem lies in its capacity to directly confront the critical challenges often associated with standalone generative models, such as reliance on noisy, unfiltered data or outdated general knowledge. Advanced multi agent and retrieval framework studies emphasize that integrating a multi phase filtering and data re ranking pipeline can significantly elevate the reliability of responses generated by artificial intelligence [16]. By forcing the language model to synthesize its assessment exclusively from the vector database populated with curated pedagogical modules and teacher rubrics, the system successfully filters out irrelevant or hallucinated information. This architectural choice is particularly crucial in educational contexts, where large language models frequently face significant challenges in evaluating higher order cognitive tasks and providing nuanced, expert level feedback without external context grounding [17]. Consequently, the SAGE platform not only scores essays with mathematical precision but also constructs diagnostic narratives that are deeply contextualized to the specific

DOI: https://doi.org/10.31004/riggs.v5i2.8979

learning objectives of the history curriculum. Once all modules ranging from class management to automated grading were fully programmed the application was compiled into a complete platform ready to enter the testing cycle.

Figure 4. System Developed Result

The testing phase is a crucial stage to ensure that the developed platform is not only free from technical errors but also has high pedagogical reliability. This systematic evaluation is consistent with the principle that empirical testing in the research and development phases is essential to validate the feasibility of the final software product [9]. The technical evaluation begins with the execution of Black Box functionality testing targeting user interaction with the system interface. The test results show that all main application features including the authentication process class management uploading of reference documents and submission of student essays are able to operate optimally. All test scenarios achieved a one hundred percent success rate without any operational failures found so the system stability is declared excellent. The evaluation then continued at the software architecture level through White Box testing involving a comprehensive review by two expert lecturers. Experts from the fields of information systems and informatics education engineering examined in detail the program code structure control logic flow database schema security and the reliability of the Gemini API integration mechanism. The results of this expert test confirmed that the system has very solid data traffic integrity and is capable of processing artificial intelligence architecture calls efficiently and securely without any parameter leaks. After the technical functionality was validated the testing phase moved to the assessment accuracy experiment conducted at SMAN 1 Blahbatuh. This test was specifically focused on measuring the effectiveness of the Retrieval Augmented Generation method in replicating the sharpness of human evaluation in the qualitative domain. This experiment involved one hundred and eighty authentic essay responses collected from thirty six eleventh grade students. The automated scores and diagnostic feedback provided by the SAGE engine were then statistically compared with manual scores from three expert history teachers using the Quadratic Weighted Kappa metric instrument. The calculation results show that the SAGE platform successfully achieved a Kappa coefficient value of 0.9133. This achievement falls into the almost perfect agreement category based on the established statistical interpretation criteria [18]. In addition to the very high level of agreement among raters this system also shows an extraordinary level of precision where 94.44 percent of the total artificial intelligence evaluations fall exactly within the ten point deviation tolerance limit from the average score of the expert teachers. A summary of all these system testing results is presented in Table 4.

Table 4. Testing Result

No	Type of Testing	Percentage or Coefficient Result	Assessment Category
1	Black Box Functionality Test	100 Percent	Excellent
2	White Box Architecture Test	Met Expert Standards	Highly Valid
3	Quadratic Weighted Kappa Accuracy Test	0.9133	Almost Perfect
4	Ten Point Tolerance Precision Test	94.44 Percent	Excellent

These highly positive statistical figures prove that context restriction through the Retrieval Augmented Generation architecture is highly effective in mitigating the risk of information hallucination in large language models. The platform does not produce biased assessments but consistently adheres to the history rubric parameters uploaded by the educator. This level of accuracy, which closely matches the objectivity of human educators leads to the

DOI: https://doi.org/10.31004/riggs.v5i2.8979

conclusion that the Smart Automated Grading Engine platform is highly feasible and ready to be widely implemented to alleviate the burden on teachers in evaluating long format academic assignments.

4. Conclusion

This research successfully developed and implemented the Smart Automated Grading Engine platform as a pioneering solution for objective and measurable automated essay assessment in the Indonesian educational context. The integration of the Retrieval Augmented Generation architecture with the Gemini API has been proven to effectively overcome the fundamental weaknesses of conventional large language models particularly in mitigating the risk of information hallucination. By implementing a strict context restriction mechanism based on teacher curated reference documents the SAGE system consistently generates evaluations and diagnostic feedback that are perfectly aligned with specific pedagogical parameters. The functional testing results confirmed that the platform possesses high operational stability and is technically ready for deployment within a dynamic digital learning environment. The empirical validation conducted at SMAN 1 Blahbatuh provides robust evidence regarding the reliability of this system in replicating the qualitative assessment sharpness of expert human educators. The achievement of a Quadratic Weighted Kappa coefficient of 0.9133 which represents an almost perfect agreement category combined with a high precision rate of 94.44 percent within a 10 point tolerance indicates that SAGE can deliver fair and accurate grading results. The primary conclusion of this study demonstrates that the application of Retrieval Augmented Generation within a Learning Management System is highly effective in enhancing administrative efficiency for teachers without compromising the academic quality of the assessment. Furthermore, the structured feedback provided by the system offers significant benefits for students in understanding their learning progress in a more transparent and timely manner. Future research and development are expected to expand the subject matter coverage and optimize the data processing speed to support a larger scale of implementation across various educational institutions and diverse academic disciplines.

Reference

[1] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.10997

[2] M. Faseeh *et al.*, "Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy," *Mathematics*, vol. 12, no. 21, Nov. 2024, doi: 10.3390/math12213416.

[3] R. Bambang, E. Saputro, and R. Hikmawan, "A Rubric-Integrated Assessment System Using a Large Language Model for Automated Essay Evaluation in Secondary Vocational Schools," *Journal of Educational Sciences*, vol. 10, no. 5, pp. 11–23, 2026, doi: 10.31258/jes.10.5.p.11-23.

[4] I. M. Y. Wirawan, I. M. Yudana, and I. N. Natajaya, "EVALUASI PELAKSANAAN LEARNING MANAGEMENT SYSTEM (LMS) DI SEKOLAH PENGGERAK SMPK 1 HARAPAN DENPASAR," *JURNAL ADMINISTRASI PENDIDIKAN INDONESIA*, vol. 13, no. 1, pp. 44–54, May 2022, doi: 10.23887/jurnal_ap.v13i1.957.

[5] P. R. Darmayasa, I. B. N. Pascima, and K. Agustini, "Pengembangan Sistem Penilaian Keaktifan Belajar Otomatis Berbasis Learning Management System (LMS) Dengan Retrieval-Augmented Generation (RAG)," *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, vol. 14, no. 3, 2025, doi: https://doi.org/10.23887/karmapati.v14i3.102876.

[6] A. A. G. P. Prameswara, I. N. I. Wiradika, L. P. E. Damayanti, and I. P. G. P. Pastika, "Development of a Large Language Model-Based Diagnostic Assessment System for Detecting Students' Initial Abilities," *Measurement in Educational Research*, vol. X, pp. 1–11, 2022, doi: 10.33292/meter.v1i1.xxx.

[7] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100417, Jun. 2025, doi: 10.1016/j.caeai.2025.100417.

[8] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," pp. 7969–7992, Oct. 2023, [Online]. Available: http://arxiv.org/abs/2305.06983

[9] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta, 2018.

[10] S. Y. Luis, D. G. Reina, and S. T. Marín, "Towards a Retrieval-Augmented Generation Framework for Originality Evaluation in Projects-Based Learning Classrooms," *Educ. Sci. (Basel).*, vol. 15, no. 6, p. 706, Jun. 2025, doi: 10.3390/educsci15060706.

[11] S. Pambudi, Herman Dwi Surjono, Totok Sukardiyono, and Akhsin Nurlayli, "A Moodle-Based Digital Learning Approach to Enhance AI Literacy Competence in Non-STEM Programs," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 15, no. 1, pp. 183–194, Mar. 2026, doi: 10.23887/janapati.v15i1.105958.

[12] I. K. Resika Arthana, N. Gunantara, M. Sudarma, and M. Sukarsa, "A Systematic Literature Review of Retrieval-Augmented Generation Implementation for Enhancing Large Language Models in Education," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 15, no. 1, pp. 91–109, Mar. 2026, doi: 10.23887/janapati.v15i1.112281.

[13] R. V. Barenji, N. Salimi, and S. Khoshgoftar, "An LLM -Powered Assessment Retrieval-Augmented Generation (RAG) For Higher Education," Jan. 2026, [Online]. Available: http://arxiv.org/abs/2601.06141

[14] N. K. Nopiani, N. Sugihartini, I. Bagus, and N. Pascima, "PENGEMBANGAN KONTEN INTERAKTIF BERBASIS MODEL DISCOVERY LEARNING PADA MATA PELAJARAN ILMU PENYAKIT DAN PENUNJANG DIAGNOSTIK KELAS XI DI SMK NEGERI 4 NEGARA," *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, vol. 11, no. 1, 2022, doi: 10.23887/karmapati.v11i1.39827.

DOI: https://doi.org/10.31004/riggs.v5i2.8979

[15] C. Pramartha, I. Koten, I. G. N. A. C. Putra, I. W. Supriana, and I. W. Arka, "Pengembangan Sistem Dokumentasi Melalui Pendekatan Ontologi untuk Praktek Budaya Bali," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 259–268, Dec. 2022, doi: 10.23887/janapati.v11i3.53939.

[16] E. Prasetio, L. B. H. Handoko, and K. Hastuti, "Improving Retrieval-Augmented Generation Performance Using the MAF-RAG Architecture, EVR–VOR Vector Retrieval, and Multi-Agent Fallback Reasoning," *Journal of Applied Informatics and Computing*, vol. 10, no. 1, pp. 212–223, Feb. 2026, doi: 10.30871/jaic.v10i1.11738.

[17] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[18] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

DOI: https://doi.org/10.31004/riggs.v5i2.8979