JUSTEK : JURNAL SAINS DAN TEKNOLOGI
http://journal.ummat.ac.id/index.php/justek
ISSN 2620-5475
Vol. 9, No. 3, September 2026, pp. 420-432

Design and Development of a Transparent AI-Driven Assessment Framework Based on BERT/RoBERTa and Explainable AI for Student Resume Evaluation

Yogi Prasetyo[1*], Yuli Sopianti[2], Latifahny Aridia Alfitri[3]
[1,2,3]Program Studi Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia, Indonesia
[1*]yogiprasetyo@upi.edu, [2]yulisopianti@upi.edu, [3]latifahny_aridia@upi.edu

ARTICLE INFO

Article History:

Received : 04-08-2026
Revised : 14-09-2026
Accepted : 15-09-2026
Online : 17-09-2026

Keywords:
Artificial Intelligence;
Explainable Artificial Intelligence;
Dual-validation Mechanism;
Intelligent Assessment;
Learning Analytics.

ABSTRACT

The evaluation of resume-based answers demands complex semantic understanding and knowledge structuring, making manual assessment prone to inter-rater inconsistency, while deep learning models such as BERT generally operate as a black box without adequate transparency. This study aims to design a Transparent AI-Driven Assessment Framework that integrates a BERT/RoBERTa model for resume assessment, an Explainable AI module based on SHAP and LIME, and a dual-validation mechanism based on multiple-choice analytics to verify learners' conceptual understanding. The study employs the Research and Development (R&D) method, covering needs analysis, architectural design, and the development of a prototype based on a microservices architecture interoperable with the Learning Management System. A formative evaluation of the prototype conducted on 15 respondents (7 lecturers and 8 students) yielded an average System Usability Scale score of 88.17 (Excellent category) and high perceived usefulness and transparency scores (4.78 and 4.72 out of 5), which preliminarily support the feasibility of the design, although these results remain indicative given the limited number of respondents.

A. INTRODUCTION

The improvement of education quality, as emphasized in Sustainable Development Goal (SDG) 4, requires assessment systems that are fair, transparent, and comprehensive. Evaluating resume- or essay-based answers is a major challenge because it requires understanding of semantics, context, and complex knowledge structures (Rani and Badusha, 2023; Chaikhamwang et al., 2025). This challenge is particularly acute in real educational contexts for instance, in undergraduate computer science courses, graders often struggle with assessing project reports where students must explain complex algorithms and design choices; in medical education, evaluating clinical reasoning essays across multiple instructors produces inconsistent quality assessments; and in large-enrollment humanities classes, providing timely, consistent feedback on research papers becomes practically infeasible. Conventional grading methods that rely on manual evaluation face significant limitations in terms of consistency, objectivity, and efficiency, particularly in large classes, which can lead to variation and grading bias that is difficult

to account for (Guskey, 2024; Malouff and Thorsteinsson, 2016). Previous studies have shown that, using identical rubrics, differences in grader characteristics can produce score disparities of up to 13 points on equivalent student work (Kates et al., 2023).

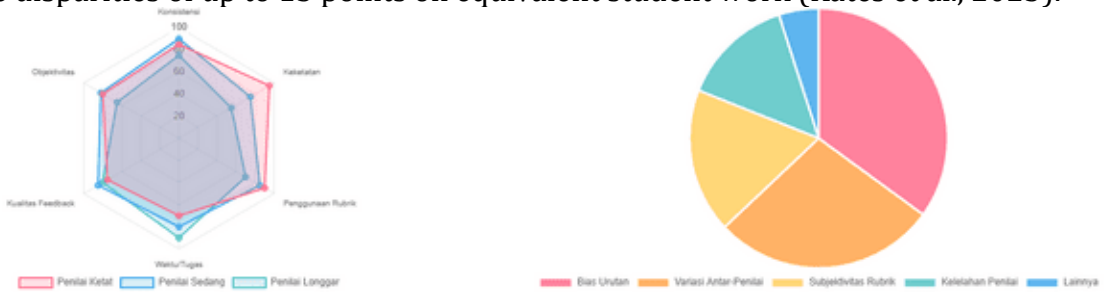

Figure 1. Illustration of Score Disparity in Manual Grading between Strict and Lenient Graders (Kates et al., 2023).

This gap has driven the adoption of artificial intelligence as an automated solution. BERT and its derivative models have demonstrated strong predictive performance in automated essay scoring (Zhong, 2024); however, most such systems still operate as black boxes without adequate transparency, raising concerns about trust, accountability, and pedagogical validity that run counter to the spirit of SDG 4 (Conijn et al., 2023). On the other hand, evaluation based solely on resumes does not guarantee that the resulting score truly reflects a learner's conceptual understanding, necessitating a complementary instrument in the form of multiple-choice test analytics as objective data on structured conceptual mastery (Rintala, 2023).

A dual-validation mechanism that integrates resume scoring with multiple-choice analytics has the potential to detect inconsistencies in understanding and generate more accurate diagnostic insights (Wanichsan et al., 2021). In addition, such an assessment system requires a scalable, interoperable, and collaborative digital platform architecture in line with SDG 17, so that it can be integrated with the Learning Management System (LMS) already used by educational institutions (Pramukastie et al., 2025; Prasetyo, 2021). Explainable AI (XAI) using SHAP and LIME has been shown to improve model transparency (Salih et al., 2025; Hermosilla et al., 2025); however, the integration of explainable AI, a dual-validation mechanism based on multiple-choice analytics, and a comprehensive digital platform architecture within a single framework remains limited in the existing literature.

To address this gap, the present study proposes the Transparent AI-Driven Assessment System (TAAS), which integrates explainable artificial intelligence, dual-validation mechanisms, and institutional interoperability into a unified framework. The research objectives are:

(1) To design and develop a transparent and explainable AI-based resume assessment model that provides interpretable scoring decisions through SHAP-based feature attribution and LIME-based local explanations, while validating its consistency against manual grading benchmarks to ensure comparability with human evaluators;

(2) To design and implement a dual-validation mechanism that integrates resume evaluation with multiple-choice analytics as a complementary assessment instrument, enabling the detection of conceptual understanding gaps through pilot implementation in real classroom settings across different disciplines;

(3) To design and deploy a scalable digital platform architecture capable of integrating assessment, validation, and analytics modules with institutional Learning Management Systems, demonstrating empirical effectiveness through controlled classroom experiments and comparative analysis of grading consistency.

This article presents the design outcomes, implementation framework, and preliminary technical validation of the TAAS as a foundation for empirical testing in subsequent research phases.

B. METHODS

This study employs a Research and Development (R&D) method aimed at producing a design and prototype of a transparent AI-based resume assessment system. The discussion in this article is limited to the design phase as a foundation for subsequent development and empirical testing. The design process was carried out through four sequential steps.

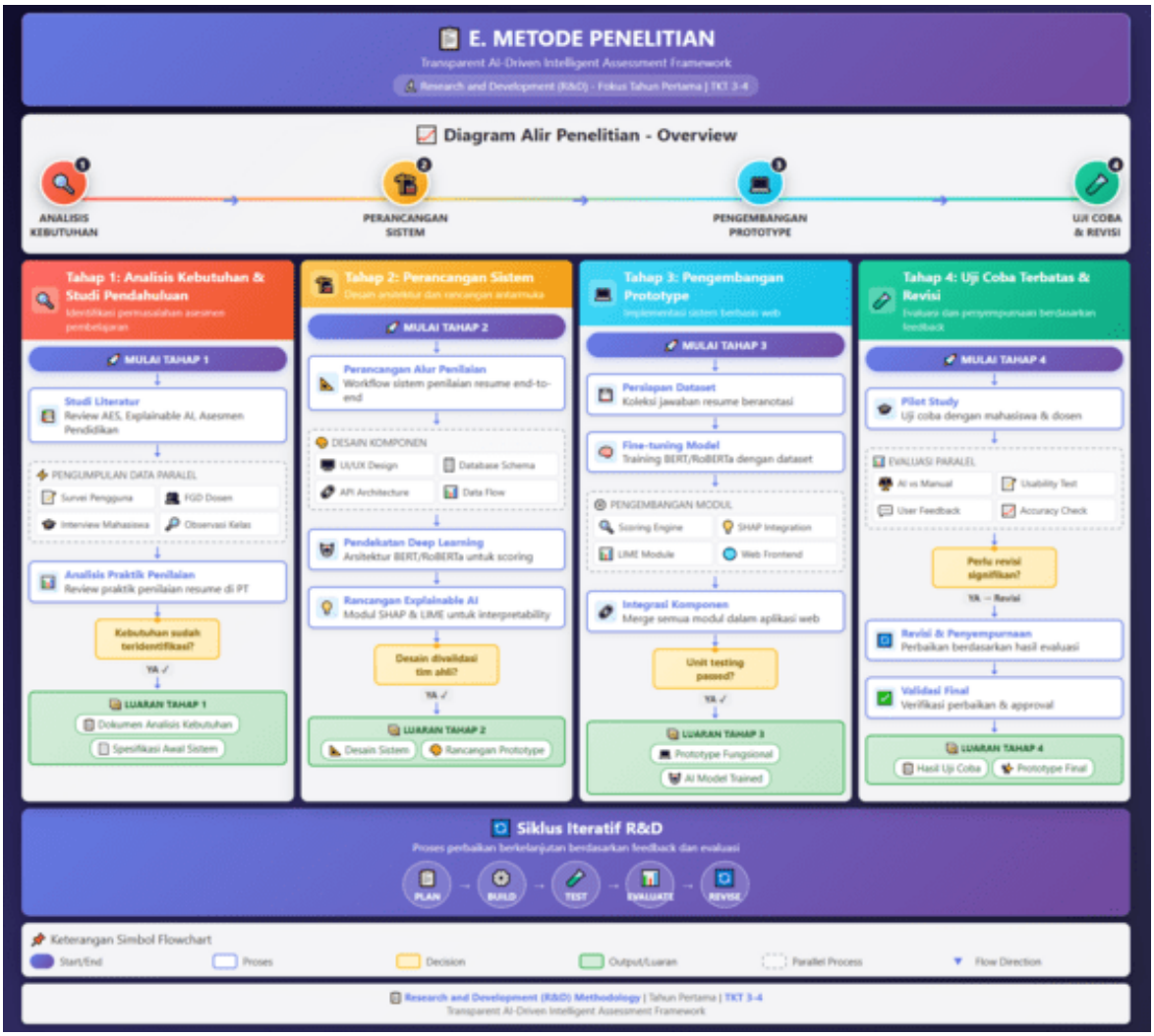

Figure 2. Research Flow Diagram of the Design Phase.

1. Needs Analysis

The first step consisted of a needs analysis, carried out through a literature review on automated essay scoring, Explainable AI, and educational assessment, as well as a Focus Group Discussion (FGD) with lecturers and students. The output of this step was a list of functional and non-functional system requirements, covering the need for assessment transparency, the need for objective validation of conceptual understanding, and the need for interoperability with the Learning Management System already used by the institution.

2. System Architecture Design

The second step was the development of the system architecture in the form of a multi-layer design consisting of three main layers: an AI-based resume assessment layer equipped with Explainable AI, a dual-validation mechanism layer that combines the resume score with multiple-choice analytics, and a digital platform layer that provides API services, an analytics dashboard, and system integration based on a microservices architecture, drawing on a Model-Driven Architecture approach for real-time business intelligence applications (Bazza et al., 2025).

3. Design of the Resume Assessment Model and Explainable AI

The third step was the design of the resume assessment model and the Explainable AI module. The resume assessment model was designed using a fine-tuning approach applied to a pre-trained BERT/RoBERTa language model with an annotated resume-answer dataset as training data, producing a numerical score representing answer quality based on predetermined assessment criteria, following a fine-tuning approach that has proven effective for automated essay scoring tasks (Zhong, 2024).

The Explainable AI module was designed using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) in a complementary manner, with output designed to be presented across three layers: feature-level (highlighting of key words or phrases), decision-level (contribution score of each assessment criterion), and pedagogical-level (actionable improvement recommendations for students), following common practice in applying SHAP and LIME to AI-based systems (Salih et al., 2025; Zhang et al., 2025).

4. Design of the Dual-Validation Mechanism and Platform Architecture

The fourth step was the design of the dual-validation mechanism and platform architecture. The dual-validation mechanism was designed to compare the resume assessment score with the results of multiple-choice test analytics analyzed using an Item Response Theory (IRT) approach, in order to detect inconsistencies between narrative construction ability and structured conceptual mastery, in line with IRT-based multiple-choice test analytics approaches (Rintala, 2023; Wanichsan et al., 2021).

The digital platform was designed using a microservices architecture with five main services: a Resume Assessment Service, an Explainability Service, a Multiple-Choice Analytics Service, a Validation Engine Service, and an Analytics Dashboard Service, each designed to be containerized and integrated with the LMS through the LTI 1.3 and OAuth 2.0 standards, referring to established practices in designing digital education platform architectures (Prasetyo, 2021; Pramukastie et al., 2025). The output of each of the four steps was reviewed internally by the research team before being realized as a functional prototype, which is planned to be tested through a limited pilot study in the next research stage.

5. User Formative Evaluation Instrument and Procedure

User evaluation was conducted as a formative usability evaluation of the prototype, given that its purpose was to capture users' initial perceptions during the design stage rather than to achieve statistical generalization. The instrument used was a condensed version combining the 10-item System Usability Scale (Brooke, 1996) to measure prototype usability with two core constructs from the Technology Acceptance Model (Davis, 1989) most relevant to the research claims, namely Perceived Usefulness (PU) and Perceived Explainability (PE), each comprising 4 items, resulting in a total instrument of 18 Likert-scale (1-5) statements.

Respondents were selected using purposive sampling, namely lecturers and students with relevant experience in resume assessment or AI-based system development, yielding 15 respondents (7 lecturers and 8 students) from seven different study programs. Each respondent tried the system prototype (20.5 minutes on average) before completing the questionnaire. Given the limited number of respondents, the analysis was conducted descriptively at both the individual-respondent and role-group levels, without reliability testing or inter-construct relationship testing, which would require a larger sample to yield stable results.

C. RESULT AND DISCUSSION

1. System Architecture

The design process produced a Transparent AI-Driven Assessment Framework architecture composed of three integrated layers: an AI-based resume assessment layer, a dual-validation layer, and a digital platform layer.

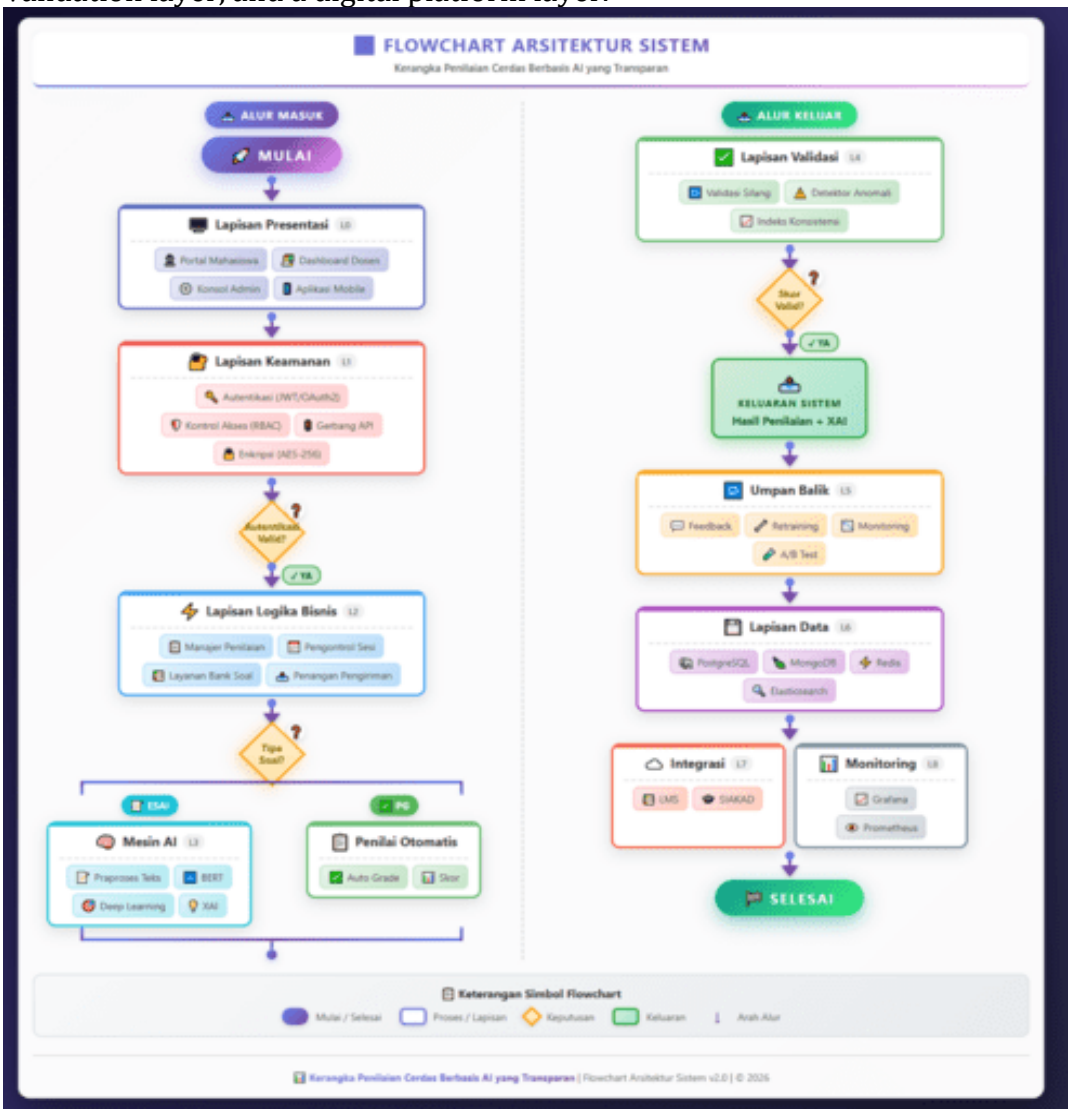

Figure 3. System Architecture of the Transparent AI-Driven Assessment Framework.

As shown in Figure 3 three layers were designed to operate modularly, facilitating development and maintenance during subsequent implementation. This modular approach also allows each component to be tested and developed separately before being fully integrated into a single assessment ecosystem.

2. Explainability Mechanism Design

The Explainable AI module was designed to present explanations at three levels simultaneously, in order to bridge the transparency needs of lecturers as assessment decision-makers and the need for actionable feedback for students.

At the feature level, the system highlights the words or key phrases in the resume text that contribute most to the score. At the decision level, the system displays the contribution score of each assessment criterion through visualizations such as waterfall plots and beeswarm plots. At the pedagogical level, the system generates concrete, actionable improvement recommendations for students.

The complementary use of SHAP and LIME was designed so that the two methods could cross-validate one another, given that they employ different computational approaches SHAP based on cooperative game theory and LIME based on a local linear model approximation yet complement each other in explaining model decisions (Hermosilla et al., 2025; Salih et al., 2025).

This three-layer design distinguishes the proposed framework from typical automated essay scoring systems, which tend to present only a final score without an explanation that can be traced back to the underlying data and assessment criteria (Emirtekin, 2025).

3. Dual-Validation Mechanism Design

The dual-validation mechanism was designed to address the limitations of resume-based evaluation alone, namely the risk that a high score does not always reflect deep conceptual understanding (Kates et al., 2023). By comparing the resume assessment score with IRT-based multiple-choice analytics results, the system design has the potential to identify patterns of inconsistency, such as students with strong narrative writing ability but weak structured conceptual mastery, or vice versa.

Such inconsistency patterns are designed to be presented as diagnostic insights for lecturers, providing a basis for more targeted learning interventions, rather than relying solely on a single score that risks concealing genuine gaps in understanding.

4. Platform Architecture and Interoperability Design

The designed microservices architecture allows each service—resume assessment, explainability, multiple-choice analytics, validation engine, and dashboard—to be scaled independently according to user load requirements, so that each component can be developed and maintained without disrupting the operation of the others.

Integration with the LMS through LTI 1.3 and OAuth 2.0 was designed so that educational institutions would not need to replace their existing learning system, but could simply connect this assessment platform as an additional service. This design is expected to support institutional scalability and cross-platform collaboration in digital education, in line with SDG 17. In addition to the backend architecture, the design stage also encompassed the design of the user interface based on human-centered design principles, so that assessment results and Explainable AI explanations could be accessed intuitively by both lecturers and students.

5. Prototype Interface Design (User Interface)

To validate the fit between the architecture in Figure 3 and user needs, the system design was realized as an interface prototype (mock-up) across three main pages: the Lecturer Dashboard, the Assessment Detail and Explainability page, and the Dual-Validation Report.

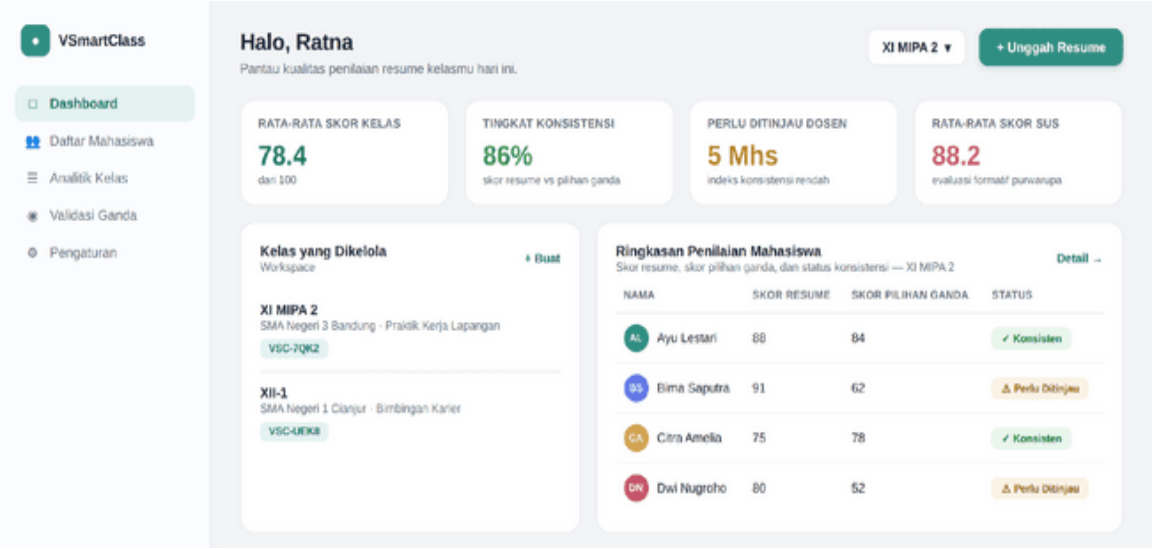

Figure 4. Lecturer Dashboard Interface Design.

The lecturer dashboard was designed to display a summary of class performance in shown in Figure 4, including the mean resume score, the consistency rate of the dual-validation results, the number of students requiring lecturer review, and the mean System Usability Scale (SUS) score from the prototype's formative evaluation. The Managed Classes panel displays the list of classes along with their class join codes, while the Student Assessment Summary panel presents the resume score, multiple-choice score, and consistency status for each student, enabling lecturers to monitor assessment progress at both the individual and class-aggregate levels within a single view.

Figure 5. Assessment Detail and Explainability Interface Design.

As shown in Figure 5, The Assessment Detail page simultaneously displays all three layers of Explainable AI explanation for a single student: a resume excerpt with key phrases highlighted according to their contribution (feature level), a chart of scores for each assessment criterion (decision level), and a box of actionable improvement recommendations (pedagogical level).

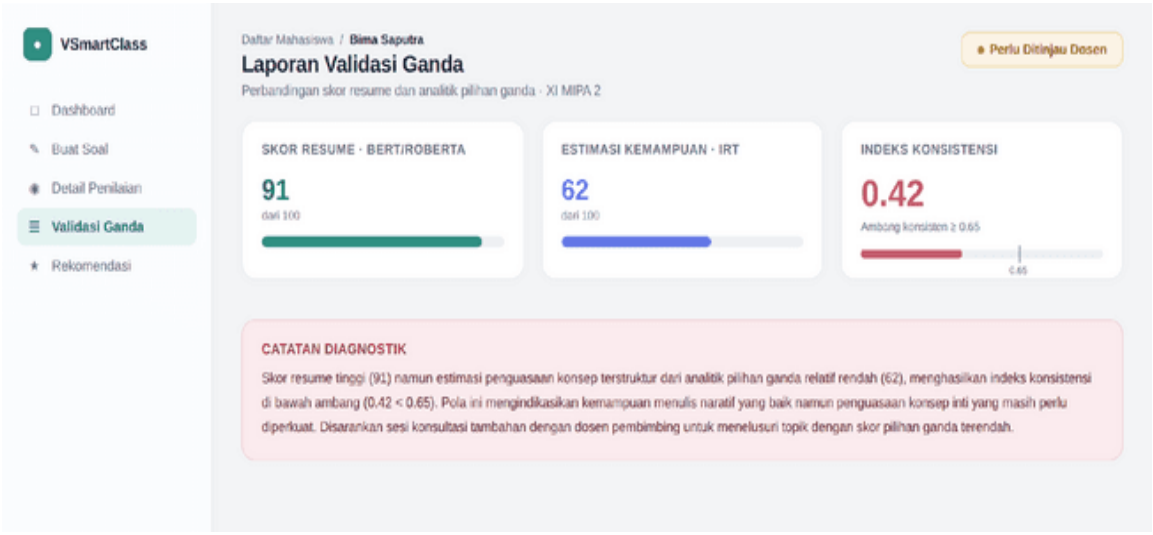

Figure 6. Dual-Validation Report Interface Design.

As shown in Figure 6 Dual-Validation Report page compares the resume score from the BERT/RoBERTa model with the estimated conceptual ability derived from IRT-based multiple-choice analytics, accompanied by a consistency index and an automated diagnostic note. This design is intended to allow lecturers to readily identify students exhibiting a pattern of strong narrative writing ability but underdeveloped structured conceptual mastery, or vice versa, without needing to read the raw data from both instruments separately.

6. Formative User Evaluation Results

The formative evaluation was conducted with 15 respondents (7 lecturers and 8 students) selected through purposive sampling from seven study programs. Table 1 presents the System Usability Scale (SUS) scores along with the mean Perceived Usefulness (PU) and Perceived Explainability (PE) scores for each respondent..

Table 1. Skor SUS, PU, dan PE per Responden

Respondent	Role	SUS	Mean PU	Mean PE
R001	Lecturer	97.5	5.00	5.00
R002	Lecturer	87.5	4.75	4.75
R003	Lecturer	85.0	4.75	4.50
R004	Lecturer	97.5	5.00	4.75
R005	Lecturer	85.0	4.75	4.75
R006	Student	85.0	4.75	4.50
R007	Student	82.5	4.50	4.50
R008	Student	92.5	5.00	5.00
R009	Student	80.0	4.25	4.25
R010	Student	87.5	5.00	5.00
R011	Student	85.0	4.75	4.75
R012	Student	85.0	4.75	4.75
R013	Lecturer	95.0	5.00	5.00
R014	Student	82.5	4.50	4.25
R015	Lecturer	95.0	5.00	5.00

As shown in Table 1, mean SUS score across all respondents was 88.17 (SD = 5.78; range 80.0-97.5), which, according to the adjective rating scale of Bangor, Kortum, and Miller (2009), falls into the "Excellent" category, well above the "Good" threshold (68) and above the typical cross-industry mean SUS score of around 68. The mean Perceived Usefulness (PU) score was 4.78 and the mean Perceived Explainability (PE) score was 4.72 on a 1-5 scale, both falling into the "Very High" category, indicating that respondents perceived the prototype as useful and found the Explainable AI explanations reasonably sensible.

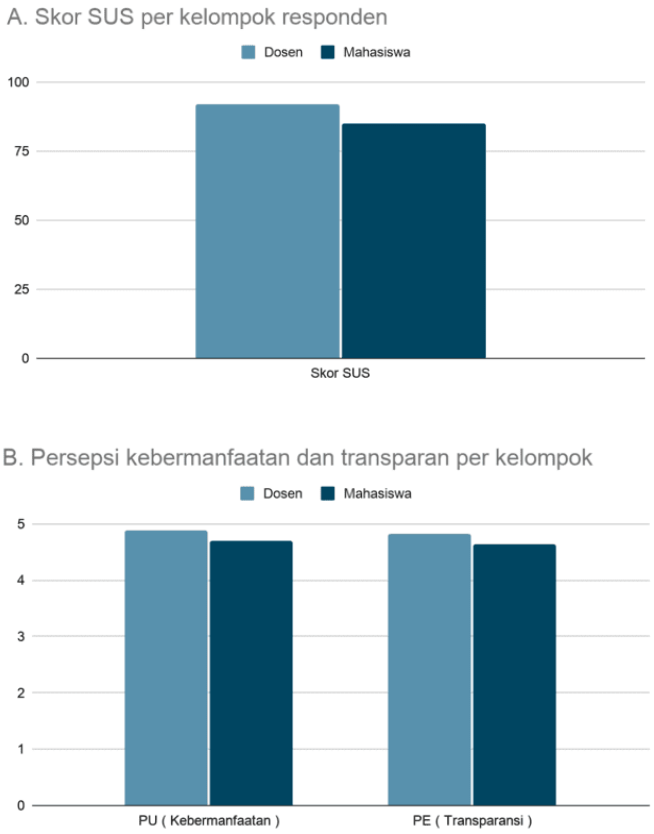

Figure 7. Comparison of SUS, PU, and PE Scores between the Lecturer and Student Groups.

As shown in Figure 7, the lecturer group exhibited slightly higher mean scores across all three metrics (SUS = 91.79; PU = 4.89; PE = 4.82) compared to the student group (SUS = 85.00; PU = 4.69; PE = 4.63). This pattern can tentatively be explained by lecturers' greater familiarity with formal rubrics and assessment criteria, making the criterion-based explanations presented by the system feel more familiar to them, whereas students may hold different interface expectations. Given the sample size in this study, this difference is descriptive in nature and has not been tested for statistical significance.

These formative evaluation results are indicative in nature given the limited number of respondents (n = 15) and the purposive sampling technique used, and therefore cannot yet be generalized to the broader population of lecturers and students. Reliability testing and statistical inter-construct relationship testing were deliberately not conducted at this stage, as they would require a larger sample to yield stable results. An evaluation involving a larger number of respondents and a random sampling design is planned for the next research stage to obtain stronger statistical validity.

7. Discussion

The proposed framework design offers an approach that differs from typical automated essay scoring systems, which tend to focus solely on predictive accuracy. The integration of explainable AI, dual validation, and an interoperable platform architecture was designed to simultaneously address three fundamental needs in educational assessment: accuracy, transparency, and scalability. The formative evaluation results, which show high SUS, PU, and PE scores, provide preliminary support for these design assumptions, although further testing on a larger sample is required. The user interface design, which combines the three explainability layers and the dual-validation report within a single view, is also intended to support adoption by lecturers without requiring them to directly interpret the technical output of the AI model. This design is consistent with previous findings that AI exhibits more stable performance on objective assessments, such as multiple-choice items, compared to subjective assessments, lending empirical justification to the use of multiple-choice analytics as validation for resume assessment (Ilieva et al., 2025; Mpolomoka, 2025).

Nevertheless, all of the claims above remain design hypotheses and still require empirical testing in real learning environments to ascertain model accuracy, the level of agreement with lecturers' assessments, and the level of user acceptance, which will constitute the agenda for the subsequent implementation and testing stage.

A critical distinction underpinning this framework's design is positioning AI as an assessment assistant rather than a replacement for lecturer expertise. The dual-validation mechanism and three layer explainability design deliberately preserve and reinforce the human decision-making loop. Specifically, the inconsistency patterns identified by the system are intended as diagnostic alerts for lecturers, not as deterministic assessment outcomes. Lecturers retain authority over assessment decisions and can exercise professional judgment when the AI flagged patterns warrant further investigation or pedagogical intervention. This is particularly important in educational contexts where assessment serves not only a summative function but also carries consequences for student progression and well-being. By surfacing the AI's reasoning through explainable outputs, the framework enables lecturers to validate the system's logic against their own professional knowledge of individual students, their learning trajectories, and contextual factors that may not be apparent in resume or multiple-choice data alone. This human-AI collaboration model aligns with emerging best practices in educational technology, which emphasize augmentation of instructor capability rather than automation of judgment (UNESCO, 2023). The high SUS and PE scores from lecturers in the formative evaluation suggest receptiveness to this collaborative approach, though the subsequent implementation phase must further examine whether lecturers perceive the system as genuinely supportive or whether concerns about over-reliance on AI assessments emerge in practice.

Beyond supporting summative assessment, the framework was designed to leverage automatic feedback generation as a mechanism for formative learning. The pedagogical level explanations of concrete, actionable improvement recommendations are generated immediately upon resume assessment, enabling rapid feedback cycles that are often infeasible with purely manual grading. This aligns with research on formative feedback showing that timely, specific, and actionable guidance is associated with improved student learning outcomes (Hattie & Timperley, 2007; Beaumont, O'Doherty, & Shannon, 2011). By providing students with criterion specific, feature-level explanations of their resume quality in addition to improvement recommendations, the system creates scaffolding for self-regulated learning and metacognitive awareness. Students are not

merely told their score but are invited to understand which aspects of their work contributed most to that score and how they might strengthen those dimensions. This transparency facilitates the feedback loop described by Carless and Boud (2018) in which students actively engage with and respond to feedback, rather than passively receiving it. Furthermore, the ability to generate immediate feedback at scale addresses a common barrier in higher education: the time constraints that often prevent lecturers from providing detailed written feedback on every assignment. Rather than replacing lecturer feedback, the automatic feedback is positioned as a first-pass formative intervention that helps students begin self-evaluation and preparation for more targeted, high-level interactions with educators. The consistency report showing where student narrative writing ability diverges from conceptual mastery may prompt lecturers to provide qualitatively different follow-up guidance than a single score would suggest, thereby enriching rather than diminishing the feedback dialogue.

D. CONCLUSION AND SUGGESTIONS

This study produced a design for a Transparent AI-Driven Assessment Framework that integrates a BERT/RoBERTa-based resume assessment model, an Explainable AI module based on SHAP and LIME, a dual-validation mechanism based on multiple-choice analytics, a microservices-based digital platform architecture interoperable with the Learning Management System, and a user interface prototype design that presents the three explainability layers and the dual-validation report in an integrated manner. The three-layer explanation design of the Explainable AI module has the potential to enhance the transparency and accountability of the assessment process, while the dual-validation mechanism has the potential to detect inconsistencies between narrative writing ability and structured conceptual mastery. A formative evaluation of the prototype with 15 respondents (7 lecturers and 8 students) showed a mean System Usability Scale score of 88.17 (Excellent category) as well as high Perceived Usefulness and Perceived Explainability scores (4.78 and 4.72 out of 5, respectively), providing preliminary support for the feasibility of the design, although this remains indicative given the limited sample size and sampling technique. Further research is needed to test this design on a larger and more representative sample, including empirically measuring the correlation between AI-generated assessment results and lecturers' manual assessments.

ACKNOWLEDGEMENT

This activity is supported by a research/community service grant from the Indonesian University of Education with contract letter number 118/UN40.F7/PT.01.01/2026 in 2026.

REFERENCES

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of Usability Studies, 4(3), 114–123.

Bazza, H., Bimonte, S., Gourti, Z., Rizzi, S., & Badir, H. (2025). An MDA approach for robotic-based real-time business intelligence applications. Data and Knowledge Engineering, 157, 102418. https://doi.org/10.1016/j.datak.2025.102418

Beaumont, C., O'Doherty, M., & Shannon, L. (2011). Reconceptualising assessment as a weapon for learning. European Journal of Engineering Education, 36(2), 137-147. https://doi.org/10.1080/03043797.2010.547941

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), Usability evaluation in industry (pp. 189–194). Taylor and Francis.

Carless, D., & Boud, D. (2018). Student voice and learning: Conceptualising capabilities for dialogic engagement. Studies in Higher Education, 43(9), 1568-1578. https://doi.org/10.1080/03075079.2016.1265196

Chaikhamwang, S., Montri, W., Fongmanee, S., & Janthajirakowit, C. (2025). Application of advanced natural language processing methodologies utilizing GPT-4o-mini to augment the efficacy of an essay examination framework on a web-based application. International Journal of Computer Applications, 6(1), 1–30. https://doi.org/10.34218/IJCA_06_01_001

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. Journal of Learning Analytics, 10(1), 37–53. https://doi.org/10.18608/jla.2023.7801

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. Applied Sciences, 15(10), 5683. https://doi.org/10.3390/app15105683

Guskey, T. R. (2024). Addressing inconsistencies in grading practices. Phi Delta Kappan, 105(8), 52–57. https://doi.org/10.1177/00317217241251883

Hattie & Timperley. (2007). Bukti bahwa timely, specific, actionable feedback meningkatkan learning outcomes

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Hermosilla, P., Berrios, S., & Allende-Cid, H. (2025). Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models. Applied Sciences, 15(13), 7329. https://doi.org/10.3390/app15137329

Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A framework for generative AI-driven assessment in higher education. Information, 16(6), 472. https://doi.org/10.3390/info16060472

Kates, S., Paulsen, T., Yntiso, S., & Tucker, J. A. (2023). Bridging the grade gap: Reducing assessment bias in a multi-grader class. Political Analysis, 31(4), 642–650. https://doi.org/10.1017/pan.2022.27

Linder, A., Nordin, M., Gerdtham, U., & Heckley, G. (2023). Grading bias and young adult mental health. Health Economics, 32(3), 675–696. https://doi.org/10.1002/hec.4639

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. Australian Journal of Education, 60(3), 245–256. https://doi.org/10.1177/0004944116664618

Mpolomoka, D. L. (2025). Utilizing artificial intelligence for assessment in higher education. Pedagogical Research, 10(3), em0243. https://doi.org/10.29333/pr/16677

Prasetyo, Y. (2021). Perencanaan arsitektur enterprise smart school menggunakan TOGAF: Studi kasus SMK Negeri 13 Bandung.

Rani, D., & Badusha, B. (2023). Intelligent descriptive answer evaluation system. Computer Science, Engineering and Technology, 1(3), 13–16. https://doi.org/10.46632/cset/1/3/3

Rintala, A. (2023). Assessment analysis: Methods and implementation options for multiple-choice exams. International Journal of Innovation in Education, 8(1), 20. https://doi.org/10.1504/IJIIE.2023.128460

Salih, A. M., et al. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), 2400304. https://doi.org/10.1002/aisy.202400304

UNESCO. (2023). Artificial intelligence and education: Guidance on generative AI in education and research. UNESCO Publishing.

Wanichsan, D., Panjaburee, P., & Chookaew, S. (2021). Enhancing knowledge integration from multiple experts to guiding personalized learning paths for testing and diagnostic systems. Computers and Education: Artificial Intelligence, 2, 100013. https://doi.org/10.1016/j.caeai.2021.100013

Zhang, G., Pan, L., Tang, F., & Yao, F. (2025). Explainable artificial intelligence in the talent recruitment process—a literature review. Cogent Business and Management, 12(1), 2570881. https://doi.org/10.1080/23311975.2025.2570881