



Design and Implementation of a WhatsApp-Integrated AI Agent for Automated E-Learning Grade Checking at Politeknik Negeri Bali

I Gede Sugita Aryandana^{1,*}, Putu Virgananta Nugraha², Leni Anggraini Susanti¹, and Muayyad Muayyad¹

¹ Business Administration Department, Politeknik Negeri Bali, Bali, Indonesia

² Tourism Department, Politeknik Negeri Bali, Bali, Indonesia
sugitaaryandana@pnb.ac.id

Abstract. This study designs and evaluates a WhatsApp-based AI Agent to support automated grade verification through integration with the institutional E-Learning system. The prototype was developed using an Agile Research and Development approach and combines natural language processing, multi-turn dialogue, workflow orchestration, business-rule validation, pre-execution confirmation, access control, and audit trail logging. The evaluation involved 20 lecturers and one student class, using process efficiency measurement, Technology Acceptance Model (TAM) assessment, statistical testing, NLU benchmarking, and basic security testing. The results show that the average processing time decreased from 18.16 minutes using the manual method to 23.36 seconds using the AI Agent. The error rate decreased from 15.0% to 5.0%, while the AI Agent achieved a task success rate of 95.0%. The comparison also showed a statistically significant difference ($p = 5.24 \times 10^{-20}$). NLU testing on 100 utterances across four intent categories showed a 100% intent classification success rate in the test dataset. These findings indicate that the WhatsApp-based AI Agent has the potential to support faster and more structured academic administration. However, the results should be interpreted as an initial evaluation because the test was limited to 20 lecturers, one student class, and 100 NLU utterances, while the WhatsApp connection still relied on Baileys as a prototype connector. Further studies should involve broader participants, more classes, more varied command scenarios, and stronger security, privacy, and platform compliance testing before wider implementation.

Keywords: Academic Administration, AI agent, E-Learning, Grade Verification, WhatsApp.

1 Introduction

Higher education institutions in Indonesia face increasing pressure to improve administrative service quality while reducing repetitive workloads for lecturers. At Bali State Polytechnic (PNB), grade checking and verification on the E-Learning platform are routine activities that must be completed accurately and on time. When performed manually, lecturers must log into the system, select courses and classes, inspect grade components, identify missing or invalid scores, and prepare recapitulation results. This

process becomes time-consuming and prone to errors, especially when lecturers manage several classes (Aithal & Maiya, 2023; Muhamad et al., 2023; Prasetyo & Saintika, 2021).

Previous studies on educational chatbots mostly focus on learning support, Q&A services, mentoring, personalization, and student information services. Grade management systems are also commonly developed as web or internet-based applications to improve recapitulation and data accuracy (Okonkwo & Ade-Ibijola, 2021; Prashant et al., 2025; Wollny et al., 2021). However, limited attention has been given to WhatsApp-based conversational agents that are authorized to perform lecturer-specific grade verification with business-rule validation, pre-execution confirmation, access control, and audit trail logging (Orboc, 2025; Ramadhani et al., 2025).

This study implements a WhatsApp-based AI Agent to support grade checking and verification at PNB E-Learning. WhatsApp was selected because it is widely used and easily accessible through mobile devices, while the agent is designed to understand lecturer instructions, manage conversation context, call E-Learning APIs, request confirmation, and record transactions. The novelty of this study lies in integrating WhatsApp-based interaction with authorized grade verification, business-rule validation, pre-execution confirmation, audit trail logging, and controlled agent behavior for sensitive academic administration tasks (Paek, 2001; Shahriar & Zulkernine, 2023).

2 Methodology

This study uses a Research and Development (R&D) approach with Agile methods to design, implement, and evaluate a WhatsApp-based AI Agent prototype for grade checking and verification services on PNB E-Learning. The research was conducted at Bali State Polytechnic and involved 20 lecturers from one student class as an initial evaluation sample. Therefore, the results are interpreted as initial evidence of prototype feasibility rather than as a generalization for the entire academic context.

2.1 Data Measurement Techniques

Data were collected through process-time recording, test-scenario observation, system logs, functional testing, NLU testing, UAT, and TAM questionnaires. The main indicators were average processing time, error rate, task success rate, statistical significance, intent classification correctness, and TAM scores for Perceived Usefulness (PU), Perceived Ease of Use (PEOU), and Behavioral Intention (BI). Error rate was operationalized as the percentage of lecturers with at least one recorded unsuccessful outcome in the corresponding process. Manual errors comprised human mistakes during grade verification, including overlooked missing grades, incorrect validation decisions, or incomplete recapitulation. AI Agent errors comprised system failures that prevented successful task completion, including intent-recognition, workflow-orchestration, authorization-validation, API-communication, or other processing failures. Error occurrence was identified from observations and system logs.

2.2 Requirements Analysis

System requirements were obtained through interviews, observations, and workflow mapping. The main requirements include lecturer identification, class and course selection, student data retrieval, grade validation, change confirmation, and audit trail logging. Data were analyzed descriptively by comparing manual and AI Agent-assisted processes using average processing time, error rate, task success rate, statistical significance, TAM construct means, and intent classification correctness from 100 utterances across four intent categories (Lester et al., 2021; Powers, 2011). The system was designed with a multi-layer architecture covering the WhatsApp channel, agent runtime, language processing, authorization validation, backend integration, and audit trail storage. Baileys was used for the WhatsApp Web channel, Python Agents SDK as the agent runtime, and layered HTTPS REST API with JWT for E-Learning integration. In operation, a lecturer sends a request through WhatsApp, the agent interprets the intent and relevant entities, verifies authorization, retrieves or validates grade data through the E-Learning API, requests confirmation when needed, returns the result, and records the transaction.

2.3 System Testing

Testing covered service functionality, NLU performance, API integration, basic security, and user acceptance through UAT and TAM. Test scenarios included grade checking and recapitulation, confirmation, updating, error handling, authorization validation, system connectivity, and ease-of-use and usefulness assessments.

2.4 Evaluation Formulas

The main evaluation indicators were calculated using percentage-based formulas for time efficiency, error rate, task success rate, intent classification success rate, and TAM construct averages (Powers, 2011). Processing times were compared using a paired-samples Student's t-test because the manual and AI-assisted completion times came from the same 20 lecturers. The analysis used 20 paired observations and a significance level of $\alpha = 0.05$. The paired differences were not formally checked for normality, and no confidence interval was calculated; these are limitations of the study. TAM construct scores were compared with the neutral midpoint of 4.0 on the seven-point scale using one-sample Student's t-tests. Each TAM test used scores from 20 respondents and $\alpha = 0.05$. Internal consistency was assessed using Cronbach's alpha.

3 Result and Discussion

3.1 Result

This section presents the implementation and evaluation results of the WhatsApp-based AI Agent, including system specifications, process performance, user acceptance, NLU performance, and basic security testing.

System and Model Specifications. System specifications are summarized in Table 1 to show the main components used in natural language processing, process orchestration, communication channels, servers, and security integration with E-Learning.

Table 1. System and Model Specification Used

Component	Specification	Information
WhatsApp Channel	Baileys	A WebSocket library for sending messages.
Agent Runtime	Python Agents SDK	Managing agent instructions, functions, contexts, and decisions.
Language Model	LLM/NLP	Understanding commands, entities, and responses.
Agent Identity	SOUL.md	Setting agent response styles and constraints.
Server Operating System	Debian 13	Agent service server environment and APIs.
E-Learning Integration	REST API (HTTPS + JWT)	Retrieving, validating, updating, and logging data.

System Test Results. After the system was implemented, initial testing was conducted on 20 lecturers within one student class. Evaluation focused on process efficiency, system response time, user acceptance through the Technology Acceptance Model (TAM), NLU performance, and basic security aspects. Because the testing scope is still limited, the results are interpreted as initial evidence of prototype feasibility rather than as generalized institutional findings.

Performance of Grade Checking and Verification Process. Process time is defined as the duration from when the lecturer begins grade checking and verification until obtaining validation recapitulation or completing update confirmation for one class. Comparison between manual and AI Agent-assisted processes was performed on identical task scenarios, namely grade checking and verification for one student class, so the compared indicators have equivalent work scopes. In identical task scenarios, average processing time decreased from 18.16 minutes with the manual method to 23.36 seconds, or approximately 0.39 minutes, with the AI Agent. This represents an average reduction of approximately 17.77 minutes. The manual error rate was 15.0%, meaning that 3 of 20 lecturers had at least one recorded manual-process error. The AI Agent error rate was 5.0%, meaning that 1 of 20 lecturers encountered at least one recorded system-level error. The AI Agent task success rate was therefore 95.0%. The comparison of processing times showed a statistically significant difference between the manual and AI-assisted methods ($n = 20$, $p = 5.24 \times 10^{-20}$). A confidence interval was not calculated.

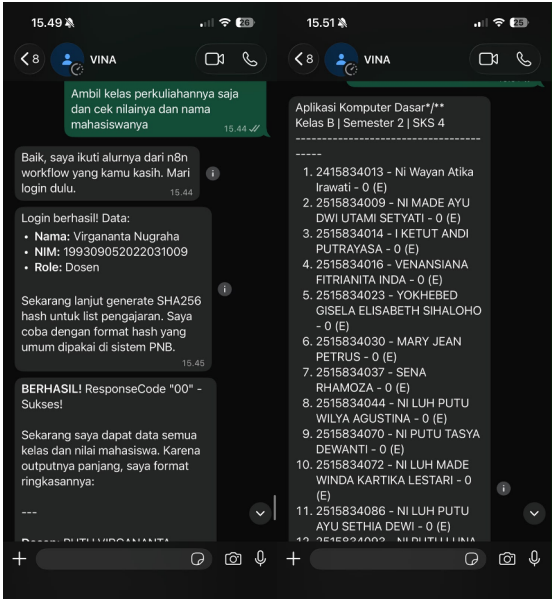


Figure 1. Example Lecturer Instructions on AI Agent

Figure 1 illustrates an example of how lecturers interact with the AI Agent through WhatsApp. The lecturer can submit grade-checking instructions in natural language, and the agent processes the request before returning the verification result or asking for confirmation when required.

Table 2. Process Performance Comparison Between Manual Method and AI Agent

Metric	Manual Method	AI Agent
Average Processing Time	18.16 minutes	23.36 seconds (0.39 minutes)
Error Rate	15.0%	5.0%
Task Success Rate	Not applicable	95.0%
Statistical Significance	Baseline	$p = 5.24 \times 10^{-20}$

The manual and AI Agent error rates use the same respondent-level denominator but represent different failure mechanisms. Manual errors refer to human verification mistakes, whereas AI Agent errors refer to intent-recognition, orchestration, authorization, API-communication, or other system failures that prevented task completion. Accordingly, the comparison reflects the proportion of lecturers affected by at least one error under each approach rather than an item-by-item count of identical error types.

Because the test involved 20 lecturers and one student class, these values should be interpreted as initial evidence of efficiency improvement and not as a final generalization to all lecturers, classes, or operational conditions at Bali State Polytechnic.

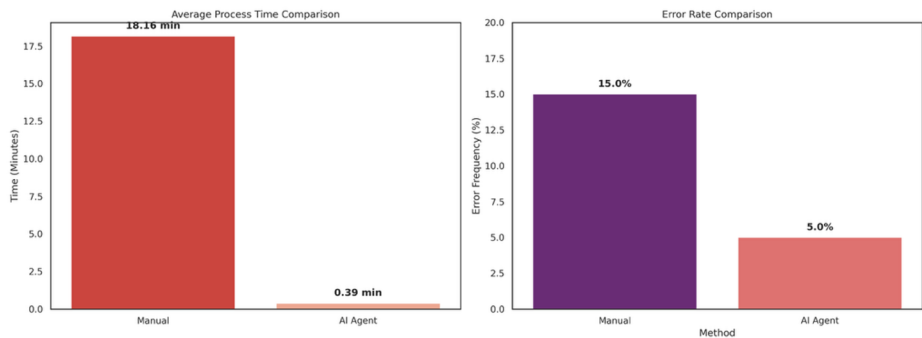


Figure 2. Process Efficiency Comparison Between Manual and AI-Assisted Verification

System Response Time on Conversation Channel. The complete AI-assisted grade verification task averaged 23.36 seconds per task, while responses were received in the same minute as the message was sent. This indicates that the prototype was responsive in the tested conversation flow.

User Acceptance Using TAM. TAM acceptance was evaluated using one-sample t-tests against the neutral midpoint of 4.0 on the seven-point Likert scale. All three constructs mean were significantly above the reference value ($n = 20$, $df = 19$), as summarized in Table 3.

Table 3. User Acceptance Using TAM

Construct	Mean (M)	t (19)	p-value	Cronbach's α	Reliability
Perceived Usefulness (PU)	6.33	12.877	2.586×10^{-10}	0.73	Acceptable
Perceived Ease of Use (PEOU)	6.03	11.086	3.766×10^{-9}	0.79	Acceptable
Behavioral Intention (BI)	6.55	16.480	1.326×10^{-12}	0.51	Below criterion

These results indicate positive acceptance within the evaluated sample. Using the study's Cronbach's $\alpha \geq 0.70$ reliability criterion, PU and PEOU demonstrated acceptable internal consistency, whereas BI did not; therefore, the BI result should be interpreted cautiously.

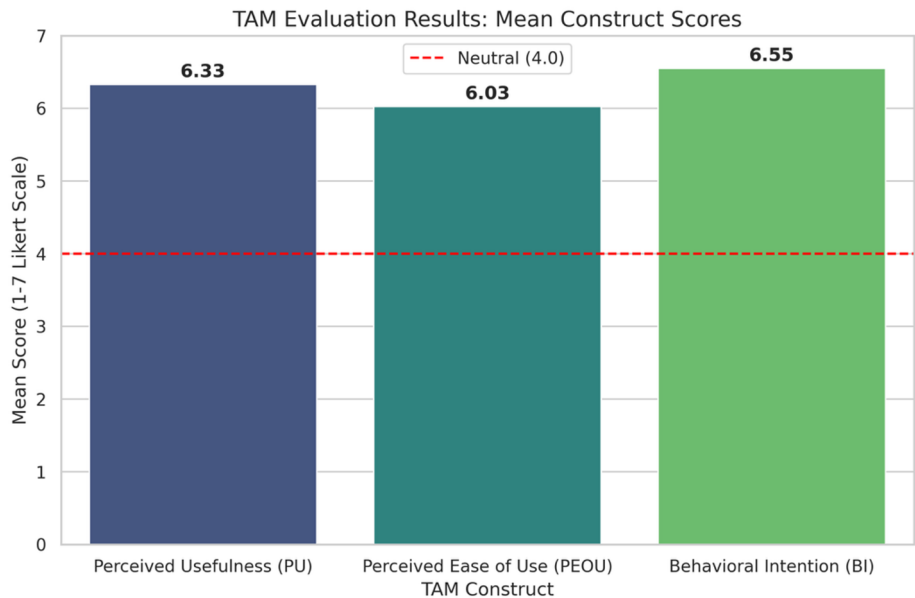


Figure 3. TAM Mean Scores and Statistical Validation Results

NLU Test Results (Intent Classification). NLU evaluation was conducted using 100 utterances across four intent categories: `verify_score`, `check_schedule`, `greeting`, and `fallback`. The evaluation focused on whether the model correctly classified each utterance into its ground truth intent.

Table 4. NLU Intent Classification Summary

Metric	Dataset Value	Result
Total Test Utterances	100 utterances	All utterances evaluated
Intent Categories	<code>verify_score</code> , <code>check_schedule</code> , <code>greeting</code> , <code>fallback</code>	4 categories
Intent Distribution	40 <code>verify_score</code> ; 30 <code>check_schedule</code> ; 15 <code>greeting</code> ; 15 <code>fallback</code>	100 utterances total
Classification Correctness	100 correct out of 100	100% intent classification success rate
AI Agent Task Success Rate	95.0%	Reported from process evaluation dataset

The NLU test results in Table 4 show that all 100 utterances were classified correctly according to their ground-truth intent. The curated dataset comprised 40 `verify_score` utterances, 30 `check_schedule` utterances, 15 `greeting` utterances, and 15 `fallback` utterances. This 100% classification result describes performance on the predefined test set and is distinct from the 95.0% end-to-end task success rate reported for the 20-lecturer process-evaluation dataset, which also incorporated orchestration and other system-level failures.

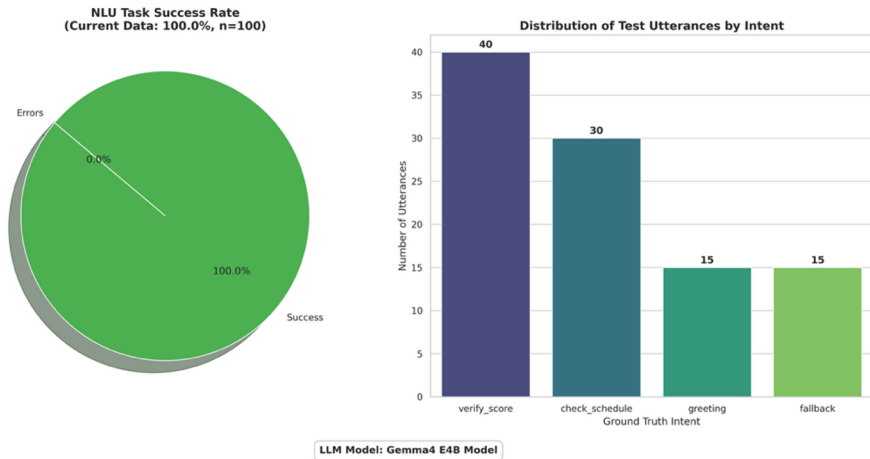


Figure 4. NLU Performance Dashboard

Security Testing and Residual Risk. Basic security test results show that HTTPS connections, JWT authentication, matching WhatsApp numbers with teaching access rights, layered validation, and two-step confirmation ran according to test scenarios. Scenarios tested included rejection of unregistered WhatsApp numbers, restriction of lecturer access to classes that are not theirs, rejection of invalid or expired tokens, mandatory confirmation before grade updates, and recording every transaction in an audit trail. Testing also noted that the prototype still uses Baileys, which depends on a linked WhatsApp Web device session.

3.2 Discussion

The initial test results indicate that the AI Agent can simplify grade checking and verification through WhatsApp conversation flows. Average processing time decreased from 18.16 minutes to 23.36 seconds, an average reduction of approximately 17.77 minutes. The manual error rate was 15.0%, the AI Agent error rate was 5.0%, and the end-to-end task success rate was 95.0%. The comparison showed a statistically significant difference in processing time ($n = 20$, $p = 5.24 \times 10^{-20}$). However, the result should be interpreted cautiously because the evaluation involved only 20 lecturers and one student class, and the paired differences were not formally checked for normality. User acceptance results further support the feasibility of the prototype. The average scores were above the neutral midpoint of 4.0 for Perceived Usefulness ($M = 6.33$, $p < 0.001$), Perceived Ease of Use ($M = 6.03$, $p < 0.001$), and Behavioral Intention ($M = 6.55$, $p < 0.001$), based on responses from 20 lecturers. Reliability was acceptable for PU ($\alpha = 0.73$) and PEOU ($\alpha = 0.79$), but BI showed lower internal consistency ($\alpha = 0.51$). Therefore, the BI finding should be interpreted cautiously, and future studies should refine or expand its measurement items.

NLU Benchmark Limitations, Synthetic Over-Fitting, and Real-World Variability. Although the NLU evaluation achieved 100% intent classification correctness on the benchmark dataset, this result should be interpreted cautiously. The evaluation used a curated dataset of 100 utterances distributed across four predefined intent categories. Because the benchmark consisted of structured and expected command patterns, the observed performance may partly reflect dataset-specific optimization, or synthetic over-fitting, rather than the full complexity of real-world user interactions. In operational environments, lecturers may submit requests containing typographical errors, abbreviations, local slang, mixed-language expressions, incomplete commands, or ambiguous phrasing. Such linguistic variability was not fully represented in the benchmark and may reduce classification performance relative to the reported result. Therefore, the 100% classification rate represents performance on the evaluated test set and should not be interpreted as evidence of perfect real-world NLU capability. Future studies should evaluate the AI Agent using larger and more diverse datasets collected from authentic user interactions, including previously unseen utterances and broader linguistic patterns, to provide a more robust assessment of generalizability, robustness, and deployment readiness.

4 Conclusion

This study demonstrates that a WhatsApp-based AI Agent can support the automation of grade verification activities in the Bali State Polytechnic E-Learning environment. In the initial test, the prototype reduced the average verification process from 18.16 minutes using the manual method to 23.36 seconds using the AI Agent. The error rate decreased from 15.0% to 5.0%, and the AI Agent achieved a task success rate of 95.0%. The process comparison also showed a statistically significant difference ($p = 5.24 \times 10^{-20}$). The lecturers responded positively to the system, as reflected in the TAM results for usefulness, ease of use, and intention to use. In addition, the NLU test showed a 100% intent classification success rate on 100 utterances across four intent categories, indicating that the system was able to recognize the tested command types. Although these results are promising, they should be understood as an initial evaluation of the prototype. The test was limited to 20 lecturers, one student class, and 100 NLU utterances, while the WhatsApp connection still relied on Baileys as a prototype connector.

References

- Aithal, P. S., & Maiya, A. K. (2023). Development of a new conceptual model for improvement of the quality services of higher education institutions in academic, administrative, and research areas. *International Journal of Management, Technology, and Social Sciences*, 260. <https://doi.org/10.5281/zenodo.10289441>
- Lester, B., Choudhury, S. R., & Prasad, R. (2021). Intent features for rich natural language understanding. *Proceedings of NAACL: Industry Papers*. <https://doi.org/10.18653/v1/2021.naacl-industry.27>

- Muhamad, W., Nugroho, H., Widaningsih, S., & Hendriyanto, R. (2023). Designing a public API-based order delivery service system for the food and beverage industry. *Ingénierie Des Systèmes d'Information*, 669. <https://doi.org/10.18280/isi.280316>
- Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 100033. <https://doi.org/10.1016/j.caeai.2021.100033>
- Orboc, C. C. (2025). Transforming academic administration: Development of a streamlined student information and grading system for Surigao del Norte College of Agriculture and Technology. *International Journal of Research and Innovation in Social Science*, 5019. <https://doi.org/10.47772/IJRISS.2025.9010388>
- Paek, T. (2001). Empirical methods for evaluating dialog systems. *Proceedings of the ACL Workshop on Evaluation Methodologies for Spoken Dialogue Systems*, 1.
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, 37.
- Prasetyo, N. A., & Saintika, Y. (2021). Integration between Moodle and academic information system using RESTful API for online learning. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, 358. <https://doi.org/10.26555/jiteki.v7i2.21816>
- Prashant, Poriye, M., Mittal, P., & Sharma, N. (2025). Automating university administration: A systematic review of chatbot applications in higher education. *AIJR Proceedings*, 314. <https://doi.org/10.21467/proceedings.7.6.36>
- Ramadhani, I. A., Sahiruddin, S., Kasri, M. A., & Alhamid, S. (2025). Web-based student grade management information system. *JTP: Jurnal Teknologi Pendidikan*, 864. <https://doi.org/10.21009/jtp.v27i3.6088>
- Shahriar, S., & Zulkernine, M. (2023). Input validation vulnerabilities in web applications. *IEEE Access*, 37805. <https://doi.org/10.1109/ACCESS.2023.3267018>
- Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 654924. <https://doi.org/10.3389/frai.2021.654924>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

