Jurnal Inovasi dan Teknologi Pembelajaran
Vol 13, No 2, (2026), page 139-157
https://doi.org/10.17977/um031v13i22026p139
P-ISSN: 2406-8780 E-ISSN: 2654-7953
Open access: https://journal-fip.um.ac.id/index.php/jinotep/index

AI-based smart web ecosystem for differentiated instruction and integrated assessment

Alfian Arif Bintara[a], Sri Wahyuni[a], Hasan Argadinata[b]

[a]*Pendidikan Bahasa Indonesia, Universitas Islam Malang*
[b] *Manajemen Pendidikan, Universitas Negeri Malang*
[a]*Jalan Mayjen Haryono No. 193, Malang, 65144 Jawa Timur, Indonesia*
[b]*Jalan Semarang 5, Malang, 65145 Jawa Timur, Indonesia*
**Corresponding author, e-mail: 22552071022@unisma.ac.id*

ARTICLE INFO

Article history:
Received: 10-08-2026
Revised: 09-10-2026
Accepted: 11-10-2026

Keywords:
Differentiated instruction; pedagogical agent; automated writing evaluation; observation report text; ecological literacy; and vocational education.

This is an open access article under the Creative Commons Attribution-ShareAlike 4.0 International license.

Copyright © 2026 by Authors. Published by Universitas Negeri Malang.

ABSTRACT

In practice, differentiated instruction is constrained by the time required to map learner profiles and the delay between submitting a draft and receiving comments. This study aimed to establish the validity and practicality of WETEL–GAIS, an AI-based smart web ecosystem, and to test its effectiveness on observation report (LHO) writing and on ecological literacy in a vocational fisheries classroom. The ecosystem's closed-domain agent, Eko-Bot, conducts Bayesian diagnostic profiling, routes students to multimodal content, provides Sentence-BERT formative scaffolding, and sets a transmediation task for summative assessment. The platform was built through the ADDIE stages and then trialled in an explanatory sequential mixed-methods study, in which the quantitative strand was a non-equivalent pretest–posttest control group quasi-experiment with 66 Year 10 students at SMKN 1 Sidayu Gresik from two purposively selected intact classes (n = 33 each). Data were analysed using Kolmogorov–Smirnov and Levene tests, independent-samples t-tests, Hake's normalised gain and Cohen's d, and triangulated with interview, observation, and telemetry data. The ecosystem proved highly valid (92.71%) and highly practical (teacher 92.50%; students 88.70%), and its profiling agreed with a standard questionnaire in 87.88% of cases. Normalised gain in writing reached 0.61 versus 0.34, $t(64) = 9.216$, $p < 0.001$, $d = 2.27$, and in ecological literacy, 0.49 versus 0.26. Integrating diagnosis, differentiation, and assessment within a single Indonesian-language platform therefore yielded measurable gains in both writing accuracy and ecological literacy.

INTRODUCTION

Writing an observation report, or laporan hasil observasi (LHO), requires students to separate statements that can be verified by observation from those that express evaluations, and to arrange the verified material according to a conventional generic structure. In a vocational fisheries programme, the requirement is more demanding than in a general school, because the object described is a technical process rather than an everyday one. A competent report on the production of fish bone meal, on salting, or on the management of washing effluent must render measurable properties of that process in language precise enough to be checked, so that the fact–opinion distinction is constitutive of the genre rather than incidental to it. Preliminary

observation in the Fishery Product Processing Agribusiness (APHPi) programme at SMKN 1 Sidayu indicated that students frequently failed to maintain that distinction, and that corrections reached them two to three days after submission, by which time they could no longer recall the reasoning behind particular sentences. Delayed assistance of this kind is associated with higher academic anxiety and with reluctance to attempt unfamiliar forms (Lee et al., 2022).

The objects available for description in a coastal vocational school also carry content that language instruction rarely exploits. Processing activities of this kind lend themselves to precise observation, yet are seldom used as ecological material in the language classroom (Sutrisna et al., 2025; Haq et al., 2025). Ecolinguistics argues that language does not merely describe environmental conditions but also shapes how they are understood and acted upon (Astawa, 2025), which suggests that a report genre may carry the environmental awareness and green skills increasingly expected of vocational graduates (Sholihah et al., 2026), and that recasting a report into another medium consolidates command of its content (Kriswanto & Fauzi, 2023).

Differentiated instruction is the standard response to such variation. Besides, because the Merdeka curriculum places learner variation at the centre of instructional planning, teachers are expected to adjust content, process, and product to differences in readiness, interest, and learning profile (Tomlinson, 2017). The supporting evidence is favourable but uneven. A meta-analysis of forty-nine studies reported a large pooled advantage for differentiated over conventional instruction (Asriadi et al., 2023). Meanwhile, a review restricted to secondary education reported a positive but considerably smaller effect with substantial heterogeneity (Smale-Jacobse et al., 2019), a discrepancy intelligible once differentiation is treated as a set of related practices rather than a single procedure. Research in Indonesia pinpoints the binding constraint: mapping learner profiles and preparing parallel materials require preparation time that teachers report they lack (Hasanah et al., 2022; Digna et al., 2024). For LHO writing, the constraint is specific, as teachers must diagnose who needs which support, prepare multiple representations of the same content, and return comments quickly enough for them to remain useful.

Artificial intelligence has been proposed as a means of relieving this constraint, and three recent reviews establish the state of the art. A systematic review of fifty-five studies of personalised adaptive learning found that only two addressed an integrated adaptive arrangement, with cross-sectional designs and self-report instruments predominating to the detriment of causal inference (Khalifeh et al., 2026). A review of thirty studies on generative artificial intelligence in K–12 education found minimal investigation of differentiated instruction and one study evaluating classroom implementation with teachers (Alfarwan, 2025). A review of 1044 studies in language learning found writing and feedback to be the largest research area. Yet, the corpus proved overwhelmingly English-medium and post-secondary, with K–12 settings accounting for 1.4% (Li et al., 2025). The quantitative syntheses agree: chatbots produce a large overall effect that is smaller in schools than in higher education (Wu & Yu, 2024). Meanwhile, automated writing evaluation produces a medium pooled effect too heterogeneous for the technology to count as one intervention type (Fleckenstein et al., 2023). Individual systems also tend to perform a single function, whether tutoring (Jančařík & Novotná, 2023), diagnostic assessment (Nabila et al., 2026), practice (Iskandar et al., 2025), or conversational rehearsal (Wicaksono, 2024).

These findings converge on one problem. What the literature has not established is whether an integrated ecosystem, in which learner diagnosis, differentiated multimodal resources, automated formative feedback, and integrated summative assessment operate as a single instructional sequence that the teacher can inspect, produces measurable gains when tested against objective pre-test and posttest measures in a school setting. Each element is separately underserved: integration itself is rare (Khalifeh et al., 2026), classroom implementation and

objective outcome measurement are rare in K–12 work (Alfarwan, 2025), and the evidence base is silent on school genres in Indonesian such as the LHO (Li et al., 2025), and teacher-facing analytic access has been examined only in isolated instances (Kim et al., 2024; Yang et al., 2025). The vocational context sharpens the question because the existing observation objects allow writing accuracy and ecological literacy to be assessed on one and the same task, an arrangement the ecopedagogical literature has so far supported descriptively rather than experimentally (Sutrisna et al., 2025; Haq et al., 2025; Sholihah et al., 2026). The present study addresses this gap by developing and testing such an ecosystem, WETEL–GAIS, in a vocational fisheries classroom.

WETEL–GAIS is a smart web ecosystem for the teaching of LHO texts whose intelligent agent, Eko-Bot, performs Bayesian diagnostic assessment, routes students to multimodal content, scaffolds drafting through closed-domain semantic feedback, and administers a transmediation task based on fishery ecology. Its claim to novelty does not rest on the co-presence of these components and is best stated by comparison with the tools schools most commonly reach for.

The first difference is domain closure with teacher curation. A general-purpose assistant such as ChatGPT generates feedback from an open model, and a controlled comparison found trained human raters outperforming it on accuracy, on the prioritisation of essential features, on clarity of direction, and on supportive tone, which the authors attributed to the model predicting text rather than understanding it (Steiss et al., 2024). Eko-Bot does not generate a judgement in that sense. It compares each sentence of a draft with a bank of 420 exemplar sentences curated by the teacher and labelled as fact or opinion, using multilingual Sentence-BERT cosine similarity against a fixed threshold (Reimers & Gurevych, 2019). The criterion is therefore explicit, inspectable, and revisable, bound to the local curriculum rather than to an open prior, and reproducible across runs. Surface-oriented aids of the Grammarly type differ in the opposite respect, addressing orthographic correctness while holding no representation of a genre-specific construct such as the fact–opinion distinction. The second difference is that diagnosis, differentiation, and assessment form a single auditable decision chain rather than three adjacent functions: the posterior probability determines the content route, formative scanning returns no mark, the summative task is set within the same platform, and every step writes to a log the teacher can read. A posterior probability can be examined, questioned, and overridden in a way a teacher's impression cannot. The third difference is that subject content is carried by the assessment itself, since the object of observation is a fishery-processing operation and both outcomes are measured on one task rather than through a separate environmental module.

The platform was designed within the framework of Technological Pedagogical Content Knowledge (Mishra & Koehler, 2006), which requires the technological component, the differentiated pedagogy, and the subject content to be developed in relation to one another. Three assumptions informed the development and the later analysis. The first concerns diagnosis: if a probabilistic classifier can infer a learning profile from a short structured dialogue, the task of mapping passes from the teacher to the system and rests on the student's responses. The second concerns the frequency of feedback: reducing the interval between submission and response from days to seconds improves no individual comment, but increases the number of complete cycles of writing, review and revision within a fixed period, each capable of carrying the three questions associated with effective feedback (Hattie & Timperley, 2007). Such cycles constitute the basic structure of self-regulated learning (Zimmerman, 2002), and acting on feedback rather than merely receiving it is a disposition acquired over time (Carless & Boud, 2018). The third concerns the affective conditions of assessment: withholding numerical scores during the formative phase is expected to position errors as information about performance rather than as judgements.

One clarification governs how the diagnostic component should be read. The study does not assume that matching instruction to a preferred sensory modality improves learning, since the meshing hypothesis has not been supported by controlled evidence (Pashler et al., 2008; Kirschner, 2017). Profiling serves the narrower purpose of establishing an entry point into the material, and all four representations remained accessible to every student throughout. The mechanism assumed is learner choice together with the management of cognitive load across complementary representations (Mayer, 2017), not modality matching.

Against this background, the study pursued four objectives: to establish the validity and practicality of the WETEL–GAIS ecosystem; to test its effectiveness on the LHO writing skill of Year 10 vocational students; to examine its effect on their ecological literacy; and to describe the pedagogical interaction with the agent, the experience of being assessed by it, and students' responses to product differentiation. Two directional hypotheses were tested, namely that writing performance and ecological literacy after the intervention would both be significantly higher in the experimental group than in the control group.

METHOD

Research design

The study was carried out in two phases because the intervention did not exist before the research began, so its effect could not be estimated until it was developed and appraised. The first phase was developmental and followed the five stages of the ADDIE model (Branch, 2009), which the study adopted because its evaluation stage is both formative and summative and therefore allows the product to be appraised before it is trialled. The second phase was evaluative, which tested the resulting platform against conventional instruction. It combined quantitative and qualitative components in an explanatory sequential arrangement, in which the quantitative results were obtained first, and the qualitative material was subsequently used to account for them (Creswell & Plano Clark, 2018). The quantitative component took the form of a quasi-experiment with a nonequivalent pretest–posttest control group design, since random assignment of individual students was not possible: class membership at the school is fixed by the timetable, and intact classes therefore had to be retained. The qualitative component intended to describe how students used the agent, how they experienced being assessed by it, and how they responded once permitted to select the medium of their final product. Table 1 sets out the design of the second phase, and Figure 1 shows the research sequence.

Table 1. Nonequivalent pretest–posttest control group design

Group	Pretest	Treatment	Posttest
Experimental (X APHPi 2)	O_1	X_1: WETEL–GAIS ecosystem	O_2
Control (X APHPi 1)	O_1	X_2: expository instruction	O_2

Note: O_1 = measurement of writing skill and ecological literacy before treatment; X_1 = learning mediated by the WETEL–GAIS ecosystem; X_2 = conventional expository learning with printed materials and uniform slides; O_2 = measurement after treatment.

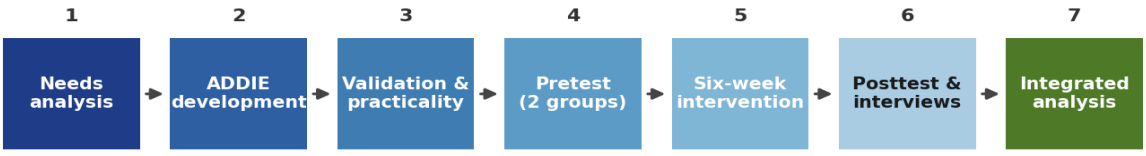


Figure 1. Sequence of research activities

Participants and setting

The research was conducted at SMKN 1 Sidayu Gresik during the even semester of the 2025/2026 academic year. The population comprised all Year 10 students in the Fishery Product Processing Agribusiness programme. Because the timetable fixes class membership, the two participating classes became the intact groups selected by purposive sampling; no random procedure operated at any level, and the design was accordingly quasi-experimental with non-equivalent intact groups. Three inclusion criteria were applied. The classes had to belong to the same programme and year level, so that curriculum and prior instruction were common to both; they had to be taught by the same Indonesian language teacher, so that teacher effects were held constant; and they had to be statistically equivalent on both outcome measures before the intervention, a criterion verified rather than assumed and reported with the pretest comparison in Table 7. Class X APHPi 2 (n = 33) served as the experimental group and learned through WETEL–GAIS, while class X APHPi 1 (n = 33) served as the control group and followed conventional expository instruction supported by printed materials and a single slide deck, giving 66 participants in total. The two groups pursued the same learning outcomes, received the same allocation of twelve face-to-face lessons and asynchronous activities over six weeks in February and March 2026, and were taught by the same teacher, so the medium of instruction and feedback was the only planned difference between them.

Intact classes carry a limitation that should be stated rather than passed over, since differences existing before the intervention cannot be excluded by the design itself. Two measures were taken in response. Baseline equivalence was tested statistically on the pretest for both outcome measures, and no inferential claim was made beyond the conditions under which that equivalence was demonstrated. Regarding research ethics, institutional permission was obtained from the school, and students and their guardians gave informed consent after the nature of participation had been explained; withdrawal was possible at any point without academic consequence, and the six interview informants are designated Student 1 to Student 6 and are not identifiable from the data reported.

Development of the WETEL–GAIS ecosystem

The platform was produced in the first phase of the study through the five stages of the ADDIE model. Each stage is reported below together with the output it yielded, so that the developmental component may be appraised independently of the experiment that followed.

The initial stage of the phase was **Analysis.** The curriculum documents for the LHO unit were examined alongside classroom observation and an interview with the class teacher, in order to establish which aspects of the genre caused the greatest difficulty and at which points the delay in feedback was most costly. Two findings governed the subsequent design: the fact–opinion distinction was the aspect students controlled least reliably, and the interval of two to three days between submission and correction was the point at which support ceased to be usable. These became the two design requirements for the platform.

The next stage was **Design.** This stage produced the flow chart of the learning sequence, the storyboard for each screen, and the content specification for each of the six channels, with the specification of the diagnostic dialogue and the weighted analytic rubric against which drafts would be assessed. Decisions referred to the TPACK framework, so the choice of technology, the differentiated pedagogy, and the LHO content were determined in relation to one another rather than in sequence.

The following stage was **Development.** This stage produced the working platform, the bank of 420 exemplar sentences curated by the teacher from a corpus of fishery observation reports and labelled as fact or opinion, and the assessment rubric in its operational form. The two decision thresholds used by the system, namely the posterior probability of 0.75 at which the diagnostic classifier returns a profile and the cosine similarity of 0.68 at which a sentence is flagged as opinion-like, were fixed at this stage and held constant thereafter. The completed platform was then appraised in January 2026 by three validators: two Indonesian teachers, who

assessed content and language, respectively, and a holder of a master's degree in educational technology, who assessed the media and system aspects.

The stage was followed by **Implementation.** This stage consisted of a six-week trial in which the experimental class worked through the platform over twelve face-to-face lessons with asynchronous activity between them, under the conditions set out above.

The last stage was **Evaluation.** Evaluation was formative and summative. The formative component drew on the written comments from the three validators before the trial and the practicality ratings from the class teacher and students at the end of the intervention. Both are reported in Table 3. The summative component drew on the pretest and posttest measures, the interview and observation material, and the platform's interaction logs.

The ecosystem comprises six channels: a home page containing the diagnostic Eko-Bot, a multimodal teaching materials channel, an assessment channel covering formative and summative tasks, a work gallery, a reflection space, and an administrator portal. The diagnostic assessment is conducted through eight dialogue items concerning activity preference. A naïve Bayes classifier estimates the posterior probability of each profile given the student's responses and returns a visual, auditory, or kinaesthetic classification once that probability exceeds 0.75; where the threshold is not reached, further items are issued automatically. Formative feedback is generated by a multilingual Sentence-BERT model (Reimers & Gurevych, 2019), which computes the cosine similarity between each sentence of a student's draft and a bank of 420 exemplar sentences labelled as fact or opinion; all were curated by the teacher from a corpus of fishery observation reports. Sentences of which similarity to an opinion exemplar exceeds 0.68 are flagged and returned with narrative advice, and no numerical score is attached. The administrator portal stores the interaction logs, which allow the teacher to monitor the class as a whole and to determine where further intervention is required (Kim et al., 2024; Yang et al., 2025).

Instructional fidelity was maintained by varying only the medium and holding the remaining conditions constant. The learning objectives, lesson sequence, summative task specification, and teacher were identical in the two classes. An implementation checklist was completed for every session in both classes, recording whether the planned activities had been delivered as designed; in the experimental class, the telemetry provided an independent record of student activity on the platform.

Instruments

Six instruments were used, and their specifications are summarised in Table 2. The writing performance test required students to compose an observation report on an authentic fishery-processing object before and after the intervention. Scripts were assessed against a weighted analytic rubric covering text structure (25%), depth of observational content (30%), accuracy of linguistic conventions (25%), and writing mechanics (20%). Each script was scored by the Eko-Bot automated evaluator and by two trained human raters, and the machine scores were retained only after their agreement with the mean of the human ratings had been verified. Ecological literacy was measured with 24 closed items on a five-point Likert scale covering ecological knowledge, environmental concern, and conservation-oriented action. The qualitative material was obtained from a participant observation sheet, a semi-structured interview guide administered to six informants purposively selected to represent all three profile categories, and the platform telemetry log, which provided an unobtrusive record of student behaviour.

Data analysis

We examined normality with the Kolmogorov–Smirnov test and homogeneity of variance with Levene's test. We then applied independent-samples t-tests to the pretest, posttest, and normalised gain. The pretest comparison also served as the baseline equivalence test. Practical magnitude was expressed as Cohen's d. Improvement was quantified using Hake's normalised gain, classified as high ($g \geq 0.70$), moderate ($0.30 \leq g < 0.70$), or low ($g < 0.30$) (Hake, 1998). We also converted Likert data on validity and practicality into feasibility percentages. Then, agreement between the classifier and the standard questionnaire was measured with Cohen's

kappa and interpreted using conventional benchmarks (Landis & Koch, 1977). All analyses were conducted at the 0.05 significance level.

Transcripts of the conversations, observation notes, and interview data were analysed by condensing the material, displaying it, and drawing conclusions from the display (Miles et al., 2014). Trustworthiness was addressed through source and method triangulation, member checking of interview transcripts with the informants concerned, and an audit trail linking each interpretation to the data on which it was based. The two components were integrated at the final stage by considering, for each statistical result, which qualitative pattern could account for it.

Table 2. Specification and psychometric quality of the research instruments

Instrument	Construct and indicators	Items	Quality evidence
Dialogic diagnostic assessment	Learning preference profile: visual, auditory, kinaesthetic	8 (adaptive)	87.88% agreement with a standard questionnaire; κ = 0.81
LHO writing performance test	Text structure; depth of observation; linguistic conventions; mechanics	1 task, 4 weighted criteria	Inter-rater r = 0.91; machine–human r = 0.87
Ecological literacy questionnaire	Ecological knowledge; environmental concern; conservation action	24	All items valid (r > 0.361); α = 0.89
Expert validation sheet	Content, linguistic, and media feasibility	3 panels	Aggregated as feasibility percentages
Practicality questionnaire	Ease of use, clarity, and workload	Teacher and student forms	Aggregated as feasibility percentages
Observation sheet, interview guide, and telemetry log	Pedagogical interaction; evaluation comfort; product choice	6 informants; 6 weeks of logs	Source and method triangulation; member checking

RESULTS

Validity and practicality of the ecosystem

Three validators appraised the ecosystem in January 2026, before it was implemented with students, and its intended users at the end of the intervention. As Table 3 shows, all components were placed in the highest category. Media validation received the highest score at 95.00%, and mean validity across the three panels was 92.71%. The single revision note came from the linguistic panel and concerned the simplification of Eko-Bot's diction, which was brought closer to an adolescent register without loss of terminological accuracy. Practicality was rated at 92.50% by the class teacher and at 88.70% by the thirty-three experimental class students.

Table 3. Expert validation and practicality of the WETEL–GAIS ecosystem

Validation panel and aspects assessed	Items	Score	%	Category
Content expert (teacher of Indonesian) *Aspects assessed: alignment with the Phase E learning outcomes; accuracy of the fact–opinion concept; relevance of the fishery-processing object to vocational learners; depth of LHO text structure; adequacy of the diagnostic item bank and the 420 exemplar sentences; sequencing of material; integration of local ecological literacy; clarity of worked examples; fit of the differentiated module; completeness of supporting multimodal material*	10	37 / 40	92.50	Highly valid

Linguistic expert (teacher of Indonesian) *Aspects assessed: conformity to Indonesian spelling conventions; effectiveness of sentence structure; accuracy of technical fishery terminology; readability for Year 10 vocational learners; consistency of terms across the six channels; clarity of Eko-Bot's diagnostic questions; register and flexibility of Eko-Bot's feedback; clarity of the audio material and reflection instructions*	8	29 / 32	90.63	Highly valid
Media expert (holder of a master's degree in educational technology) *Aspects assessed: interface layout across devices; legibility, contrast and visual design; navigation across the six channels; system stability under simultaneous access; audio-visual quality of the video and infographic material; integration of Eko-Bot in the drafting page; clarity of the administrator portal; interactivity of the work gallery; fit of the media formats to differentiation; data security and draft storage*	10	38 / 40	95.00	Highly valid
Mean validity	**28**		**92.71**	**Highly valid**
Practicality: class teacher *Aspects assessed: class management and draft monitoring; clarity of the diagnostic output; efficiency of feedback; ease of running product differentiation; clarity of the teacher guidance; fit of the assessment instrument to curriculum criteria; readiness of the material bank; support for integrating local ecological literacy; ease of grading final work; overall reduction of administrative load*	10	37 / 40	92.50	Highly practical
Practicality: students (n = 33) *Aspects assessed: access and operation; appeal and legibility of the display; speed and intelligibility of Eko-Bot's help; usefulness of its advice for distinguishing fact from opinion; relevance of the local ecological material; freedom to choose the final product; clarity of the on-screen instructions; the work gallery; working without a numerical mark; overall preference for the approach*	10	35.48 / 40	88.70	Highly practical

Note: Every item was rated on a four-point scale, so the maximum score is four times the number of items, and the percentage is the score obtained divided by that maximum. Mean validity is the arithmetic mean of the three panel percentages. The three panels appraised the platform in January 2026, before the trial; the practicality questionnaires were administered at the end of the intervention in March 2026. One revision note was recorded by the linguistic panel and is described below; the content and media panels recorded none and declared the platform ready for trial without substantial revision.

Learner profiling and content differentiation

The intervention began with the dialogic diagnostic on the platform's home page. As shown in Figure 2, Eko-Bot first establishes the student's identity and then presents preference items in everyday language, processing each response through the classifier until the confidence threshold is exceeded. All 33 students in the experimental class completed the diagnostic in a single session, with a mean duration of 6.4 minutes, and none required assistance from the teacher.

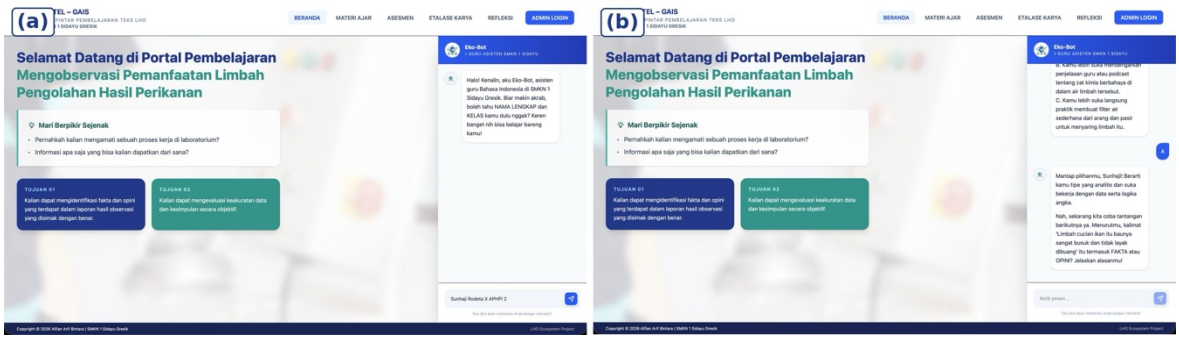


Figure 2. Dialogue-based diagnostic assessment on the WETEL–GAIS home page: (a) identity initiation by Eko-Bot; (b) activity preference items used for Bayesian inference

Note: The interface is in Indonesian. The navigation bar reads Beranda (Home), Materi Ajar (Teaching Materials), Asesmen (Assessment), Etalase Karya (Work Gallery), Refleksi (Reflection), and Admin Login. The banner reads 'Welcome to the learning portal for observing the use of fishery processing waste'. The prompt box Mari Berpikir Sejenak ('Let us think for a moment') asks 'Have you ever observed a work process in a laboratory?' and 'What information could you obtain from it?' The two objective cards read 'Objective 1: you can correctly identify facts and opinions contained in the observation report you have attended to' and 'Objective 2: you can evaluate the accuracy of data and conclusions objectively'. In (a), Eko-Bot opens with 'Hello, I am Eko-Bot, the Indonesian language teaching assistant at SMKN 1 Sidayu Gresik. May I know your full name and class first?' In (b), it continues, 'Your choice indicates that you are the analytical type who likes working with data and numerical logic. Now for the next challenge: in your view, is the sentence "The smell of fish washing waste is very foul, and it is unfit to be discarded" a FACT or an OPINION? Explain your reasoning.'

Table 4 displays the final distribution. Visual profiles were most common (42.42%), followed by kinaesthetic (30.30%) and auditory (27.27%) profiles, with an average posterior probability of 0.86. To evaluate the precision of classification, a different, common questionnaire about learning preferences was given independently and contrasted with the agent's outcome. The two instruments agreed for 29 of the 33 students, or 87.88%, resulting in a Cohen's kappa of 0.81.

Table 4. Distribution of learning preference profiles in the experimental class as classified by Eko-Bot (n=33)

Classification	f	%	Mean posterior probability	Primary content route
Visual	14	42.42	0.88	Slides and infographics
Auditory	9	27.27	0.85	Podcast
Kinaesthetic	10	30.30	0.84	Practical video and observation project
Total	33	100.00	0.86	-

The diagnostic also served as conceptual apperception. Figure 3 illustrates how Eko-Bot examines students' initial sensitivity to the fact–opinion distinction using contextual sentences about effluent from the fish-washing tanks, and then provides corrective feedback that explains the linguistic reasoning behind the judgement, instead of stating only whether the answer was correct.

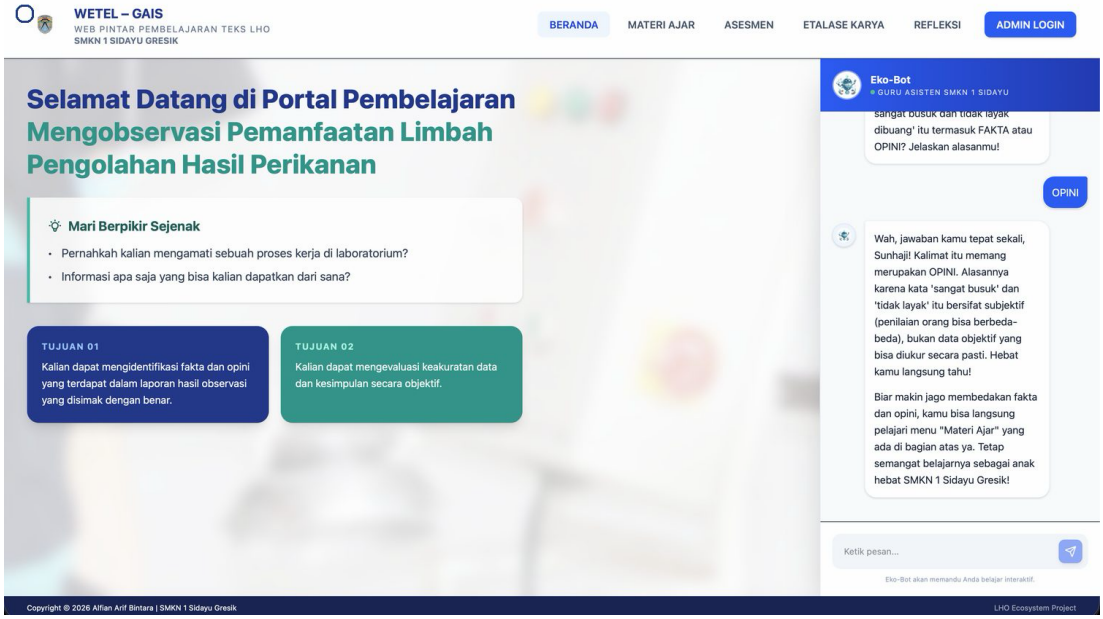


Figure 3. Corrective feedback from Eko-Bot during the fact opinion discrimination exercise on the home page

Note: Translation of the exchange: The student answers “OPINI (OPINION)”; Eko-Bot replies, ‘Your answer is exactly right. The sentence is indeed an OPINION because “very foul” and “unfit” are subjective, since people may judge them differently, rather than objective data that can be measured precisely. Well spotted. To get better at telling fact from opinion, go to the Materi Ajar (Teaching Materials) menu at the top. Keep up the good work as a student of SMKN 1 Sidayu Gresik.’

Students were subsequently routed to the teaching materials channel, which provides four content-equivalent modes: a contextual instructional video on fish waste utilisation, a structured slide deck, an infographic, and a 23-minute podcast (Figure 4). Over the six weeks, the telemetry recorded 722 material visits, of which 78.6% were consistent with the recommended route. Since students were not prevented from opening any of the other modes, this proportion indicates that the recommendation was generally followed by choice.

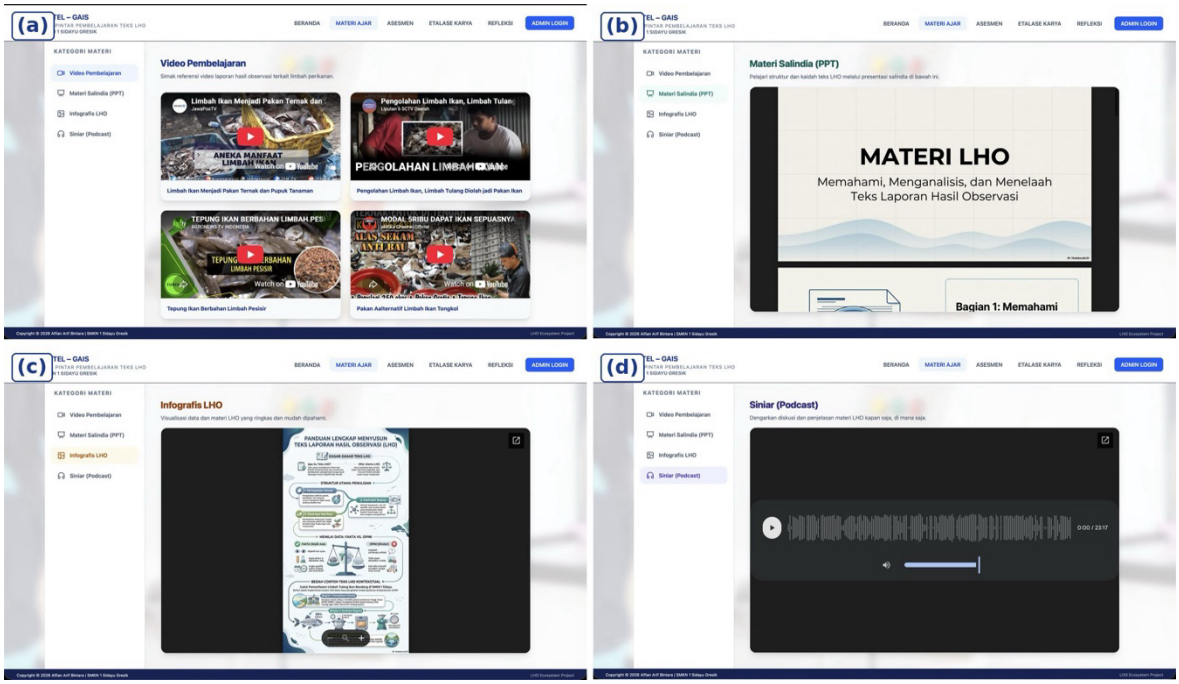


Figure 4. Multimodal content differentiation in the teaching materials channel: (a) instructional video; (b) slide deck; (c) LHO infographic; (d) podcast

Note: The sidebar heading Kategori Materi (Material Categories) lists Video Pembelajaran (Instructional Video), Materi Salindia (Slide Deck), Infografis LHO (LHO Infographic), and Siniar (Podcast). The slide in (b) reads ‘LHO material: understanding, analysing and examining observation report texts’, and the infographic in (c) is headed ‘A complete guide to composing observation report texts’.

Effectiveness on LHO writing skill

Process differentiation operated through the formative assessment channel (Figure 5). Students typed or uploaded a draft, which Eko-Bot examined sentence by sentence, flagging those identified as subjective opinion and attaching narrative advice without a numerical mark. The median feedback latency was 3.8 seconds, so that the writing, review, and revision cycle could be repeated several times within a single session. Students in the experimental class produced an average of 2.7 draft versions per task, compared with 1.2 in the control class, where drafts remained with the teacher for two to three days before corrections were returned.

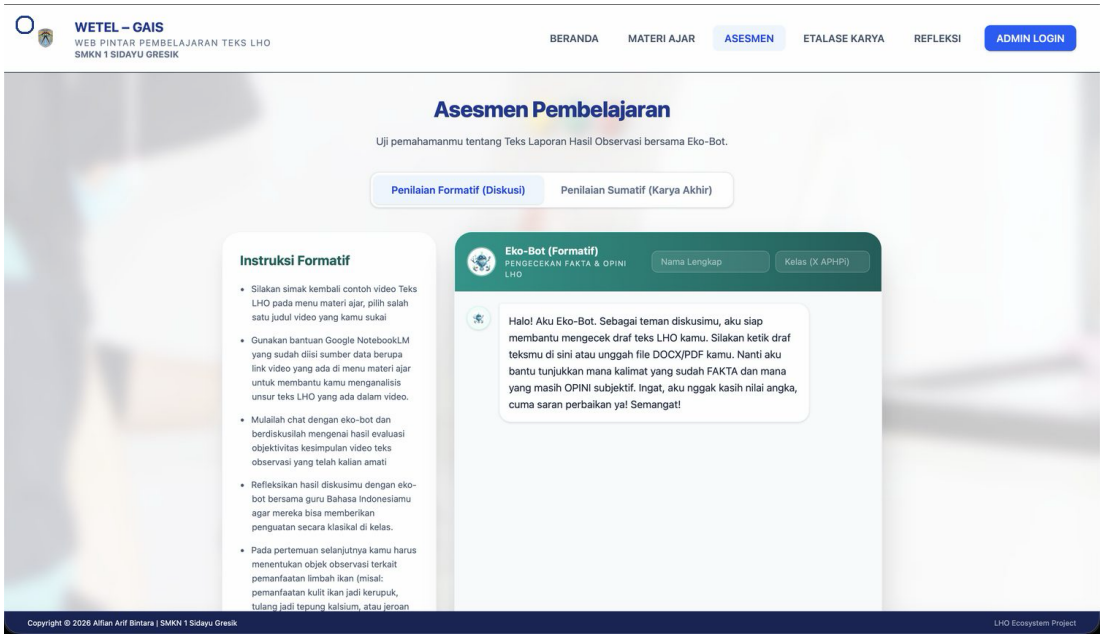


Figure 5. Formative assessment interface: fact–opinion scanning of an LHO draft by Eko-Bot

Note: The page is headed Asesmen Pembelajaran (Learning Assessment), subtitled 'Test your understanding of observation report texts with Eko-Bot', and offers the tabs Penilaian Formatif (Diskusi) (Formative Assessment – Discussion) and Penilaian Sumatif (Karya Akhir) (Summative Assessment – Final Work). The left-hand panel is headed Instruksi Formatif (Formative Instructions). Eko-Bot opens with 'Hello, I am Eko-Bot. As your discussion partner, I am ready to help check your LHO draft. Type your draft here or upload your DOCX or PDF file. I will then point out which sentences are already FACT and which are still subjective OPINION. Remember, I do not give numerical marks, only suggestions for improvement.'

Descriptive statistics for the writing performance test are reported in Table 5. The two classes began from comparable positions, with pretest means of 58.91 and 57.15, both below the mastery criterion. After six weeks, the experimental mean had risen to 83.79 with a narrower dispersion, while the control mean reached 71.76. Posttest scores in the experimental class ranged from 73 to 91, which places the lowest experimental score above the lowest score recorded in the control class.

Table 5. Descriptive statistics for LHO writing scores

Group	n	Pretest M	SD	Posttest M	SD
Experimental (X APHPi 2)	33	58.91	5.32	83.79	4.79
Control (X APHPi 1)	33	57.15	7.00	71.76	5.77

The assumption checks reported in Table 6 confirmed that parametric analysis was appropriate. All four distributions were normal, with Kolmogorov–Smirnov significance values between 0.562 and 0.879, and the variances were homogeneous at both the pretest and posttest.

Table 6. Kolmogorov–Smirnov normality and Levene homogeneity test results

Test	Data	Statistic	Sig.	Remarks
Normality	Experimental pretest	0.113	0.748	Normal
Normality	Experimental posttest	0.110	0.778	Normal
Normality	Control pretest	0.098	0.879	Normal
Normality	Control posttest	0.133	0.562	Normal
Homogeneity	Pretest (Levene F)	2.314	0.133	Homogeneous
Homogeneity	Posttest (Levene F)	2.054	0.157	Homogeneous

Table 7 summarises the hypothesis tests. The pretest difference between the groups was not significant, t(64) = 1.149, p = 0.255, which establishes baseline equivalence. The posttest difference was significant, t(64) = 9.216, p < 0.001, with Cohen's d = 2.27, as was the difference in normalised gain, t(64) = 10.624, p < 0.001, d = 2.62. Figure 6 displays the change in mean scores for both groups.

Table 7. Independent-samples t-test results and effect sizes

Comparison	t	df	Sig. (2-tailed)	Cohen's d
Pretest, experimental–control	1.149	64	0.255	0.28
Posttest, experimental–control	9.216	64	< 0.001	2.27
Normalised gain, experimental–control	10.624	64	< 0.001	2.62

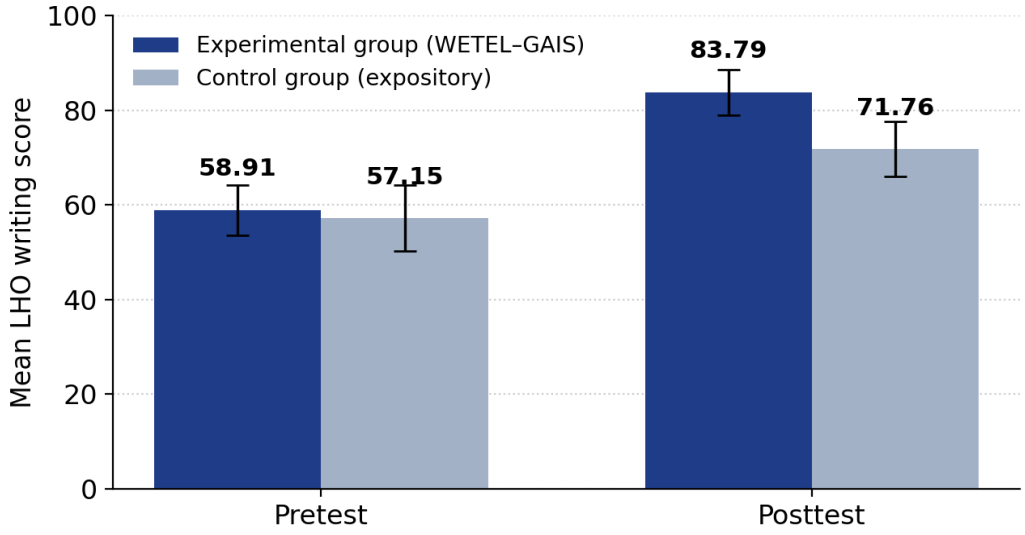


Figure 6. Mean pretest and posttest LHO writing scores for both groups (error bars show standard deviations)

The normalised gain analysis in Table 8 provides information that the means alone do not convey. The experimental class achieved a mean gain of 0.61, at the upper end of the moderate band, with 18.18% of students in the high category and none in the low category. The control class achieved 0.34, and 36.36% of its students remained in the low category.

Table 8. Normalised gain in writing skill and distribution across categories

Group	Mean N-gain	Category	High f (%)	Moderate f (%)	Low f (%)
Experimental	0.61	Moderate	6 (18.18)	27 (81.82)	0 (0.00)
Control	0.34	Moderate	0 (0.00)	21 (63.64)	12 (36.36)

When the gains are decomposed by rubric aspect (Figure 7), the advantage of the experimental class is present in all four dimensions but is not evenly distributed. The largest gains occurred in text structure (0.67 vs. 0.39) and in depth of observational content (0.63 vs. 0.36). Linguistic conventions reached 0.57 vs. 0.30, and writing mechanics showed the smallest gain (0.53 vs. 0.28). The control class therefore remained at or below the boundary of the moderate band on both aspects. The validity of the automatic scoring was also examined. Eko-Bot's scores correlated with the mean of the two human raters (r = 0.87, p < 0.001), with a mean absolute difference of 3.4 points on a 100-point scale, while the correlation between the two human raters was r = 0.91.

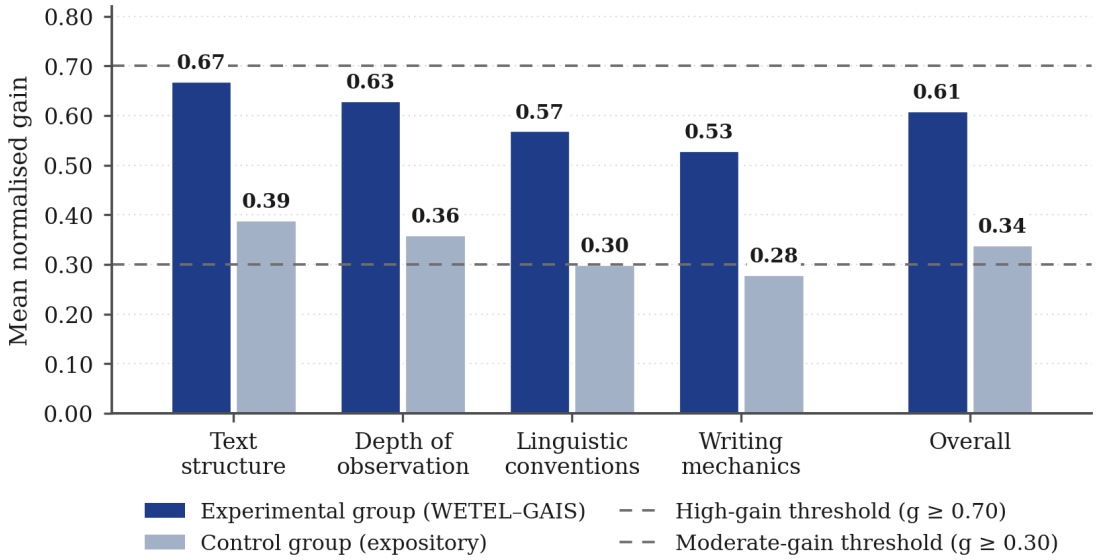


Figure 7. Mean normalised gain by writing assessment aspect for both groups

Impact on ecological literacy

Ecological literacy became the second outcome measured, using the 24-item questionnaire across three indicators. As Table 9 shows, both groups improved, but at different rates. The normalised gain of 0.49 in the experimental class falls within the moderate band, compared with 0.26 in the control class, and the difference after the intervention was significant, $t(64) = 7.182$, $p < 0.001$, $d = 1.77$. Examined by indicator, the advantage of the experimental class was greatest for ecological knowledge (0.55 vs. 0.31), followed by environmental concern (0.50 vs. 0.26), and smallest for conservation-oriented action (0.42 vs. 0.21).

Table 9. Ecological literacy scores before and after the intervention

Group	Pre M (SD)	Post M (SD)	N-gain	Category
Experimental	64.15 (7.34)	82.25 (5.20)	0.49	Moderate
Control	61.33 (3.92)	71.49 (6.87)	0.26	Low

Pedagogical interaction, psychological comfort and transmediation

The qualitative component provides an account of the processes underlying these figures. Over the six weeks, the telemetry recorded 1,247 formative messages from students to Eko-Bot, an average of 37.8 per student, and 81.82% of students uploaded at least two drafts. Condensation of the conversation data produced three recurring interaction patterns. Requests for conceptual clarification accounted for 38.2% of exchanges and most often concerned the difference between describing the parts of an object and describing its benefits. Requests for contextual examples accounted for 27.4%. Negotiation of draft revisions, in which a student proposed a corrected sentence and asked for it to be scanned again, accounted for 34.4%. The third pattern was most closely associated with high attainment, consistent with its structure, since it involves setting a target, monitoring the feedback received, and adjusting the strategy accordingly.

The six interviews yielded consistent accounts of the experience of being assessed. Student 1 reported feeling relaxed while being corrected by Eko-Bot and unafraid of making mistakes, since no mark was returned and only suggestions were given. Student 2 reported that the agent addressed the difficulty of starting a text, because questions could be asked and answered immediately, and the draft was, therefore, completed more quickly. Student 3 described a change in her own writing criteria (seeing how many of her sentences contained opinion, she replaced them with figures obtained from her own weighing in the laboratory). The class teacher reported a corresponding change in her role, noting that she at the time could move around the

classroom and support the groups that required attention, since the mechanical correction of drafts was being handled by the system.

Product differentiation was implemented as a transmediation task in the summative assessment. Eleven students (33.33%) desired an infographic; nine (27.27%) selected a journalistic vlog; seven (21.21%) favored a podcast; and six (18.18%) preferred a blog article. All works were assessed using a holistic rubric, and then displayed in the work gallery. The choice of medium generally corresponded to the profile identified at the diagnostic stage. Student 4 reported that the transmediation process had deepened her command of the content rather than merely varying the task, since producing a vlog from her own report required her to recall the report's content.

DISCUSSION

One constraint of the design should be stated before the results are interpreted, because it governs the strength of every claim that follows. The experiment compared the ecosystem as a whole against conventional instruction and did not vary its components independently. Nothing in the design therefore isolates the contribution of diagnostic routing, of immediate semantic scaffolding or of the withholding of marks from one another. Where the three assumptions set out earlier are examined below, the telemetry, interview, and observation data indicate whether an observed pattern is consistent with a proposed mechanism; they do not experimentally establish that mechanism. The wording used throughout this section is intended to preserve that distinction.

The results can be interpreted in the three assumptions set out earlier, which appear to have operated in combination rather than independently. The first concerns diagnosis: it concluded that manual differentiation has been reported to burden teachers, particularly at the stage of initial mapping (Digna et al., 2024; Hasanah et al., 2022), and in the present study, that stage was transferred to the system without a loss of accuracy, since agreement with a conventional questionnaire fell within the almost perfect band on the usual benchmarks (Landis & Koch, 1977). The significance of this lies less in the confirmation of a profile than in the fact that the routing decision becomes open to inspection, because a posterior probability can be examined, questioned, and overridden in a way that a teacher's impression cannot. To that extent, the study expands the existing case for chatbots as diagnostic instruments (Nabila et al., 2026) by adding transparency to feasibility.

The second assumption offers the clearest account of the difference between the groups, and it concerns how often a student passes through the cycle of writing, reviewing, and revising rather than the quality of any single comment. When the interval between submission and response falls from days to seconds, the cost of testing a formulation falls with it, and a student uncertain whether a sentence states a fact or an evaluation can settle the question while the intention behind the sentence is still available to them. Revision then forms part of composing rather than a separate episode of correction recalled with difficulty several days afterwards. The draft counts recorded by the platform are consistent with this account, since within the same six weeks and with the same material available, the experimental class produced 2.7 draft versions per task vs. 1.2 in the control class. Each additional cycle can carry feedback on the direction of learning, the progress achieved, and the action required next (Hattie & Timperley, 2007), so what plausibly differs between the conditions is the cumulative quantity of guided practice rather than the sophistication of any individual response. The telemetry gives the process an observable form. The interaction pattern most closely associated with attainment was the negotiation of revisions rather than the request for explanation, which corresponds to the self-regulatory cycle of goal setting, monitoring, and adjustment (Zimmerman, 2002), and the students' tendency to respond to feedback by proposing alternatives rather than by accepting it as given suggests that feedback literacy was developing during the intervention (Carless & Boud, 2018). These are associations observed within a single condition rather than effects separated by the design. Immediate semantic scaffolding may accordingly have increased the opportunities for guided practice, but the present study cannot show that this increase was its distinct contribution. Comparable dynamics have been reported for intelligent agents in mathematics

(Laja et al., 2025; Jančařík & Novotná, 2023), and the present findings are consistent with their extension to writing instruction in Indonesian.

The third assumption concerns the affective conditions of assessment, and two established accounts converge on why a response carrying no number may be received differently from one that does. The affective filter hypothesis holds that anxiety, low self-confidence, and low motivation raise a barrier that prevents available input from being used for acquisition, so the same material yields less when the learner is defensive than when the learner is at ease (Krashen, 1982). A mark attached to a draft is an evaluative act, and in a class where writing is already the least secure skill, it invites the comparison and the self-judgement that raise that barrier, whereas a narrative comment identifying which sentence states an evaluation rather than a fact does not. Self-determination theory supplies the complementary account. Feedback may be informational, supporting the need for competence by indicating what has been achieved and what remains, or controlling, shifting the perceived locus of causality outward and undermining the autonomy on which intrinsic motivation depends (Ryan & Deci, 2000). A score returned after every scan belongs to the second category, since it converts the exercise into a performance to be judged, while a diagnostic comment that leaves the revision to the student belongs to the first. The pattern in the present data is consistent with both accounts. No student in the experimental class fell into the low-gain category, against 36.36% in the control class, and the interview material indicates that the weakest writers continued to revise rather than withdrawing after an unfavourable judgement, which accords with evidence that non-judgemental conversational agents reduce the affective filter and encourage linguistic risk-taking (Lee et al., 2022). The outcome should be attributed to a design decision rather than to artificial intelligence as such, since a system configured to return a mark after every scan would in all probability have produced a different affective result. That proposition was not tested here, and establishing it would require a comparison of scored and unscored formative conditions within the same platform.\

The effect sizes obtained require cautious interpretation, and the discrepancy with the pooled estimates in the wider literature is better addressed by asking which features of the comparison differ rather than by recording that the present figure is the larger one. Automated writing evaluation yields a medium pooled effect with substantial heterogeneity (Fleckenstein et al., 2023), and chatbot interventions yield smaller effects in schools than in higher education and smaller effects in extended than in brief implementations (Wu & Yu, 2024); a Cohen's d of 2.27 stands well above both. Several differences plausibly account for part of that gap. The comparison condition here was expository teaching with manual correction returned after two to three days, a less demanding standard than the technology-supported or feedback-rich controls common in the meta-analytic corpora, and an effect size is a function of the control as much as of the treatment. The duration was six weeks, short enough for the same meta-analytic evidence to predict an inflated estimate, since brief interventions outperform longer ones partly because novelty has not yet decayed (Wu & Yu, 2024).

The participants were Year 10 vocational students rather than the undergraduates who dominate that corpus, and the genre was unfamiliar to them, so the room for improvement was correspondingly greater. The outcome was a rubric-scored performance on one school genre rather than a general proficiency measure or a self-report scale, and narrow outcomes closely aligned to what was taught ordinarily yield larger effects than broad ones. The feedback modality was closed-domain and criterion-referenced against a curated exemplar bank rather than open generative text, which removes a source of variability present in studies using general-purpose models. The measurement approach shared that alignment, since the analytic rubric against which scripts were scored covered the same constructs the platform was built to teach, and instruments constructed independently of an intervention generally return more conservative estimates. Implementation was intensive and fully delivered, with a completed fidelity checklist for every session in both classes, whereas pooled estimates average across implementations of uneven completeness. Finally, two intact classes at a single school were compared, so unmeasured differences at the class level cannot be excluded. On balance, the direction of the finding is more securely established than its magnitude, and the figure should be

read as the effect of this ecosystem against this comparison condition under these conditions rather than as an estimate of what AI-supported writing instruction achieves in general.

A second qualification concerns the diagnostic component, and it is one that reviewers in this field increasingly raise. The high level of agreement between the classifier and the standard questionnaire demonstrates consistency of measurement, but it does not demonstrate that routing students according to preference produced the learning gain. Because all four representations remained available to every student and no crossover design was used, the contribution of profile-based routing cannot be separated from the contribution of providing four well-constructed representations of the same content. Given that the meshing hypothesis has repeatedly failed to obtain empirical support (Pashler et al., 2008; Kirschner, 2017), the more parsimonious interpretation attributes the benefit to representational variety and learner choice. The telemetry finding that 78.6% of visits followed a recommendation that was never enforced is consistent with this interpretation, as is the expectation that complementary representations reduce the cognitive load imposed by an unfamiliar genre (Mayer, 2017). Resolving the question would require a comparison of matched and random routing within the same platform.

The ecological literacy results provide experimental support for a body of ecopedagogical scholarship that has so far been largely based on descriptive and qualitative work (Sutrisna et al., 2025; Haq et al., 2025; Sholihah et al., 2026). They can also be understood in ecolinguistic terms. The proposition that language shapes how the natural world is treated (Astawa, 2025) acquires a concrete mechanism when the requirements of the genre oblige students to record the calcium content of milkfish bone meal or to state the boiling temperature of processed waste precisely. Under those requirements, waste is reconstituted as material with measurable properties rather than as an object of aversion. The gradient across the three indicators is consistent with this interpretation, since knowledge improved most and conservation-oriented action least, which is what a six-week language intervention rather than a behaviour-change programme might reasonably be expected to achieve.

For the Indonesian literature on differentiated instruction, which has largely examined manually enacted strategies (Rachmadhani & Kamalia, 2023; Fahriani & Karkono, 2024; Sari, 2024), these findings indicate that automation need not deprive differentiation of its humane character. On the evidence obtained, the teacher's role changed rather than diminished, moving from the mass correction of drafts towards decisions about which students required attention, informed by the interaction data available in the portal (Kim et al., 2024; Yang et al., 2025). In comparison with the use of open generative models for language differentiation (Wicaksono, 2024), the closed-domain architecture adopted here represents a defensible compromise for school settings, since feedback remains pedagogically substantial while staying bounded by a bank of exemplars that the teacher has curated and can therefore inspect and revise. The transmediation task extends the argument that recasting a report into another medium consolidates understanding (Kriswanto & Fauzi, 2023) by adding a form of digital publication that supplies a genuine audience.

Several limitations define the boundary within which these findings hold, and are stated here so the conclusions are not extended past that boundary. The quasi-experimental design with two intact classes at a single school restricts generalisation, and random assignment at the level of the individual student was not possible. Six weeks is too short a period to establish durability, and no delayed posttest was administered. Part of the effect may be attributable to novelty. The automated evaluation addressed the fact–opinion distinction and generic structure, and did not address stylistic subtlety, cohesion across paragraphs, or the rhetorical quality of description. The exemplar bank of 420 sentences, although curated, is small enough that its coverage may influence which sentences are flagged, and the two operating thresholds were fixed rather than tuned against a held-out set. Retention, transfer to other genres, and possible over-reliance on automated support were not measured. Replication across several sites with random assignment, delayed measurement, an expanded exemplar bank, and an explicit test of the routing assumption would address most of these limitations. The addition of ecopedagogically framed game elements, which have been proposed as a means of sustaining

engagement over longer periods (Fadhilah, 2025), would allow the question of durability to be examined directly.

CONCLUSION

This study set out to pursue four objectives, and each is addressed in turn below. The ecosystem was judged valid by three validators (two teachers of Indonesian, for content and language, and a specialist in educational technology, for the media and system aspects) and practical by both the teacher and the students who used it. Students taught through it improved their observation report writing substantially more than those taught by conventional expository methods, and the advantage fell in generic structure and the depth of observational content, the two aspects the system targets, rather than in surface mechanics; no student in the experimental class remained in the low-gain category, so the benefit reached the weakest writers as well. Ecological literacy also improved more under the intervention, graded across its three indicators, with ecological knowledge responding most strongly, environmental concern next and conservation-oriented action least, which places the contribution of the system at the cognitive rather than the behavioural end of that construct. The qualitative evidence indicates that trustworthy diagnostic profiling, immediate feedback carrying no evaluative judgement, and freedom to choose the medium of the final product together reduced anxiety about assessment and sustained independent revision, while the teacher's role moved from correcting drafts in bulk towards deciding which students required attention.

The principal contribution is not that an artificial intelligence system can assist the teaching of writing, which is established, but that learner diagnosis, differentiated provision, formative feedback, and summative assessment can operate as one inspectable sequence in a school language other than English, and that a closed feedback domain curated by the teacher suffices without recourse to an open generative model. The conclusions hold under stated conditions: two intact classes at one school over six weeks without delayed measurement, so the direction of the effect is better established than its magnitude, and routing by learning preference cannot be separated from the provision of several representations among which students could choose. The agent is, accordingly, best positioned as a formative partner and the interaction data as a basis for locating where teacher intervention is needed. Extension to other genres, a larger exemplar bank, and multi-site randomised designs with delayed posttests are required before any strong claim of effectiveness.

Acknowledgments

The authors are grateful to SMKN 1 Sidayu Gresik, and in particular to the principal, the Indonesian language teaching staff and the Year 10 students of the Fishery Product Processing Agribusiness programme, for the permission, support, and cooperation that made this research possible.

Author contributions

AAB designed the study, developed the platform, collected and analysed the data, and drafted the manuscript. SW and HA contributed to the instructional design, the validation procedure and the critical revision of the manuscript. All authors have read and approved the final version.

Funding

This research received no funding or specific grant from any governmental, private or non-profit organisation.

Conflict of interest

The authors declare that there is no conflict of interest in this study.

Data availability statement

The data supporting the findings of this study are held by the corresponding author and can be made available on reasonable request.

REFERENCES

Alfarwan, A. (2025). Generative AI use in K-12 education: A systematic review. *Frontiers in Education, 10*, 1647573. https://doi.org/10.3389/feduc.2025.1647573

Asriadi, M. A., Hadi, S., Istiyono, E., & Retnawati, H. (2023). Does differentiated instruction affect learning outcome? Systematic review and meta-analysis. *Journal of Pedagogical Research, 7*(5), 18–33. https://doi.org/10.33902/JPR.202322021

Astawa, I. N. (2025). Exploring ecolinguistics: A new paradigm for tourism language learning. *Englisia: Journal of Language, Education, and Humanities, 12*(2). https://doi.org/10.22373/ej.v12i2.27648

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. https://doi.org/10.1007/978-0-387-09506-6

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Digna, D., Minsih, M., & Widyasari, C. (2024). Teachers' perceptions of differentiated learning in the independent curriculum. *Indonesian Journal of Elementary Education, 7*(2). https://doi.org/10.23887/ijee.v7i2.54770

Fadhilah, M. N. (2025). Pemanfaatan gamifikasi dalam ekopedagogi untuk meningkatkan literasi anak. *Ghancaran: Jurnal Pendidikan Bahasa dan Sastra Indonesia*. https://doi.org/10.19105/ghancaran.vi.17198

Fahriani, D. K., & Karkono, K. (2024). Dampak strategi pembelajaran berdiferensiasi terhadap hasil belajar menulis surat. *Journal of Language Literature and Arts, 4*(11), 1110–1116. https://doi.org/10.17977/um064v4i112024p1110-1116

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74. https://doi.org/10.1119/1.18809

Haq, E. D., Masdul, M. R., & Larasati, O. (2025). Ecopedagogy based curriculum innovation to foster environmental sustainability awareness among Indonesian Generation Z. *Khazanah: Journal of Islamic Education and Science, 1*(1), 17–31. https://doi.org/10.61815/khazanah.v1i1.863

Hasanah, E., Suyatno, S., Maryani, I., Al Badar, M. I., Fitria, Y., & Patmasari, L. (2022). Conceptual model of differentiated-instruction (DI) based on teachers' experiences in Indonesia. *Education Sciences, 12*(10), 650. https://doi.org/10.3390/educsci12100650

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

Iskandar, A. A., Megasari, R., Nazir, S., & Riza, L. S. (2025). Development of a drill-and-practice chatbot for enhancing English pronunciation through interactive dialogue exercises. *Jurnal Paedagogy, 12*(4), 1185–1197. https://doi.org/10.33394/jp.v12i4.17760

Jančařík, A., & Novotná, J. (2023). Using AI chatbot for tutoring and differentiated learning scaffolding. *Journal of Education Culture and Society, 14*(2), 285–296. https://doi.org/10.15503/jecs2023.2.285.296

Khalifeh, F., Santiago, R., & Palau, R. (2026). Redefining personalized learning in the artificial intelligence era: An updated systematic review from 2019 to 2025. *Smart Learning Environments, 13*, 19. https://doi.org/10.1186/s40561-026-00440-6

Kim, M., Kim, S., & Lee, S. (2024). Designing prompt analytics dashboards to analyze student–ChatGPT interactions in EFL writing. *arXiv*. https://doi.org/10.48550/arXiv.2405.19691

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education, 106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press. https://www.sdkrashen.com/content/books/principles_and_practice.pdf

Kriswanto, M., & Fauzi, N. B. (2023). Inovasi diferensiasi produk dengan metode alih wahana pada materi teks laporan hasil observasi. *Diglosia: Jurnal Kajian Bahasa, Sastra, dan Pengajarannya*. https://diglosiaunmul.com/index.php/diglosia/article/view/535

Laja, Y. P. W., Delvion, E. B. S., & Siahaan, M. M. L. (2025). Chatbot artificial intelligence (AI) to supporting differentiated mathematics learning. *Jurnal Pendidikan MIPA, 26*(1), 118–130. https://doi.org/10.23960/jpmipa.v26i1.pp118-130

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, Y., Hwang, G., & Chen, P. (2022). Impacts of an AI-based chatbot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational Technology Research and Development, 70*, 1843–1865. https://doi.org/10.1007/s11423-022-10142-8

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence, 9*, 100445. https://doi.org/10.1016/j.caeai.2025.100445

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning, 33*(5), 403–423. https://doi.org/10.1111/jcal.12197

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Nabila, A. S., Prastikawati, E. F., & Lestari, M. Y. W. (2026). Leveraging Edbot.AI as a diagnostic assessment tool to enhance the students' English achievement. *Linguistic, English Education and Art (LEEA) Journal, 9*(2), 261–281. https://journal.ipm2kpe.or.id/index.php/LEEA/article/view/19088

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*(3), 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x

Rachmadhani, S. A. D., & Kamalia, P. U. (2023). The effect of differentiated learning strategies on student learning outcomes. *Asatiza: Jurnal Pendidikan, 4*(3). https://doi.org/10.46963/asatiza.v4i3.1231

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sari, S. P. (2024). Penerapan strategi pembelajaran berdiferensiasi dalam pembelajaran teks biografi di SMA. *Pembahsi: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 14*(2). https://doi.org/10.31851/pembahsi.v14i2.16304

Sholihah, F., Herlambang, Y. T., & Yuniarti, Y. (2026). Ecopedagogy for equitable green skills development in school: A systematic review. *Journal of Inclusive Education and Social Welfare, 3*(2). https://doi.org/10.70211/wesw.v3i2.637

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology, 10*, 2366. https://doi.org/10.3389/fpsyg.2019.02366

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Sutrisna, I. P. G., Bhandesa, A. M., Wulansari, N. T., & Susanta, I. P. A. E. (2025). Ecopedagogic-based education to foster ecological awareness in Indonesian language learning. *Utamax: Journal of Ultimate Research and Trends in Education, 7*(2), 124–135. https://doi.org/10.31849/8jrnyn11

Tomlinson, C. A. (2017). *How to differentiate instruction in academically diverse classrooms* (3rd ed.). ASCD.

Wicaksono, A. (2024). Differentiated approach in speaking skills learning through ChatGPT. *Tarling: Journal of Language Education, 10*(1). https://doi.org/10.24090/tarling.v10i1.16181

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students' learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology, 55*(1), 10–33. https://doi.org/10.1111/bjet.13334

Yang, C., Wang, D., & Chen, G. (2025). Chat-LAD: Enhancing teacher understanding of learning analytics dashboard with AI-empowered explanations. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale* (pp. 197–201). Association for Computing Machinery. https://doi.org/10.1145/3698205.3733922

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2