Vol. 7, No. 1, 2026, pp. 1790-1807
DOI: https://doi.org/10.29210/994020

Contents lists available at Elora Center

Lentera Negeri

Journal homepage: https://lentera.eloracenter.org/lentera

Effectiveness of an AI-based learning assistant as a medium for strengthening students' higher order thinking skills (HOTS)

Nafra Kristia Fitri*), Muhammad Anwar, Titi Sriwahyuni, Fahmi Rizal
Postgraduate Technical and Vocational Education Program, Universitas Negeri Padang, Padang, Indonesia.

Article Info

Article history:
Received Aug 17th, 2026
Revised Sep 02th, 2026
Accepted Sep 21th, 2026

Keyword:
AI-based learning assistant,
Chatgpt,
Higher order thinking skills,
Time series design,
Vocational education.

ABSTRACT

This study examines the use of an AI-based learning assistant (ChatGPT) and changes in students' Higher-Order Thinking Skills (HOTS) in the Wired and Wireless Networking course within the Computer and Networking Engineering (TKJ) program. HOTS encompasses the abilities to analyze (C4), evaluate (C5), and create (C6) based on Bloom's revised taxonomy. The study employed a quantitative approach with a one-group time-series pre-experimental design, involving 15 11th-grade students selected through purposive sampling. The researchers collected data through problem-based HOTS essay tests, questionnaires, and observations. The instruments showed content validity with Aiken's V values ranging from 0.838 to 0.852, and the HOTS test showed reliability with Cronbach's $\alpha = 0.934$. The Lilliefors test indicated that the data were normally distributed across the three sessions. The mean HOTS score increased from 66.41 in Session 1 to 72.88 in Session 2 and 77.56 in Session 3. A repeated-measures ANOVA analysis revealed a significant difference, $F(2,28) = 54.993$, $p < 0.001$, $\eta p^2 = 0.797$, with a large effect size. The Greenhouse–Geisser correction still yielded significant results. The overall N-Gain value of 0.335 fell into the moderate category. At the same time, student acceptance of the AI-Based Learning Assistant reached 80.00%, which is classified as good. The findings indicate an increase in HOTS during AI-assisted learning.

Corresponding Author:
Nafra Kristia Fitri,
nafrakristiafitri1401@gmail.com

Introduction

Vocational High Schools are designed to prepare graduates who are work-ready, adaptive to technological change, and professionally competent in specific occupational fields. Accordingly, vocational education requires a balance between theoretical understanding and the development of practical and cognitive competencies (Taali et al., 2024). In the context of Industry 4.0 and increasingly complex workplace demands, Higher-Order Thinking Skills (HOTS), particularly students' abilities to analyze information, evaluate alternatives, and solve complex problems, have become increasingly important for vocational learners (Habib et al., 2025). These competencies are particularly relevant in Computer and Network Engineering, where students are expected not only to recall technical procedures but also to diagnose network problems, interpret technical information, evaluate alternative solutions, and make appropriate decisions.

Preliminary academic records from Grade XI students at SMK Muhammadiyah 1 Padang indicated that students' average scores remained within a relatively satisfactory range during the 2024–2026 academic years.

As presented in Table 1, the mean final scores were 78 in 2024, decreased to 76 in 2025, and increased to 79 in 2026. Although these descriptive data indicate year-to-year variation, they should not be interpreted as evidence of a statistically significant decline or improvement because the available school-level means do not provide sufficient student-level observations for a valid inferential trend analysis. Therefore, the three-year academic records are used in this study only as contextual evidence of the learning situation rather than as direct evidence of students' HOTS deficits.

Table 1. Average Student Scores of Grade XI over the Last Three Years

Year	Mean Mid-Term	Mean UTS	Mean UAS	Mean Final Score
2024	78	76	80	78
2025	76	75	78	76
2026	80	78	79	79

Because general academic achievement does not directly measure higher-order cognition, the present study did not use these school examination scores as the dependent measure of HOTS. Instead, HOTS was operationally defined as students' demonstrated ability to analyze technical problems (C4), evaluate alternative solutions (C5), and create or design appropriate solutions (C6) within the Cabled and Wireless Network Technology subject. The HOTS construct was measured through a contextual, problem-based essay test developed from the Learning Outcomes and Learning Objectives (Tujuan Pembelajaran/TP) of the subject. The assessment required students to analyze network-installation and maintenance problems, evaluate technical alternatives, and design appropriate installation, maintenance, and testing procedures. This operationalization provides a direct measure of higher-order cognitive performance rather than inferring HOTS from general academic grades. The instrument structure and scoring rubric are detailed in the Method section.

The quality of the HOTS instrument was examined through both content validation and empirical item analysis. Five experts evaluated the instrument, resulting in an overall Aiken's V of 0.852, with the assessed dimensions covering content, construction, language, item fit, and the scoring rubric. The instrument was subsequently piloted with 15 Grade XII students who were not included in the main study. Of the initial 60 items, 37 items met the empirical validity criterion, whereas 23 items were excluded or recommended for replacement based on item validity and/or discrimination results. The retained instrument demonstrated high internal consistency, with Cronbach's Alpha of 0.934 overall; the alpha coefficients for the three topic groups were 0.905, 0.829, and 0.883, respectively.

The preliminary classroom investigation further indicated that Problem-Based Learning (PBL) had been implemented in the TKJ learning environment, but the integration of contemporary digital technologies remained limited. Based on teacher interviews, Canva and Google Forms were used as supporting digital tools, whereas ChatGPT had not previously been systematically integrated into classroom learning. This condition provided the instructional context for examining the use of an AI-Based Learning Assistant as an additional learning support. Importantly, the present study does not assume that AI use automatically improves HOTS. Rather, it examines whether students' HOTS scores demonstrate a systematic pattern of improvement during repeated measurements conducted throughout AI-assisted learning.

Previous studies have reported both potential benefits and limitations of AI-assisted learning for higher-order thinking. Aulia et al. (2019) reported that HOTS-oriented assessment in vocational education requires stronger attention to analysis and evaluation. Habib, Desky, Anwar, et al. (2026) reported that AI can facilitate access to information and support comprehension, while emphasizing the need for appropriate pedagogical guidance. Desky et al. (2025) found that ChatGPT used as a virtual tutor could support students' HOTS, particularly analysis and problem solving. Similarly, Habib, Desky, Jalinus, et al. (2026) reported potential contributions of ChatGPT to critical, analytical, and ethical thinking. Other studies have also reported positive associations or learning gains associated with AI-supported learning (Dedy et al., 2025; Habib, Desky, & Yulastri, 2026). However, these findings do not establish that AI-based learning assistance produces the same effects across different vocational subjects and instructional contexts.

A methodological gap therefore remains concerning how HOTS changes across repeated learning measurements when an AI-based learning assistant is integrated into a specific vocational competency. Many intervention studies rely on a single pretest-posttest comparison, which provides limited information about the pattern of change during the learning process. The present study addresses this issue by using a one-group time-series design involving repeated HOTS measurements across learning sessions. The repeated

measurements make it possible to examine whether students' HOTS scores show a consistent pattern of change rather than relying solely on a difference between two measurement points.

The present study was conducted as a single-school exploratory intervention study involving Grade XI students in the TKJ program at SMK Muhammadiyah 1 Padang. The sample consisted of 15 students from one intact class selected purposively because the class met the instructional and technological requirements of the study. The study therefore does not claim population-wide generalizability or direct superiority over conventional instruction. Instead, it provides within-group evidence concerning changes in HOTS following the integration of an AI-Based Learning Assistant in Cabled and Wireless Network Technology learning. The sampling frame, sample characteristics, and intervention procedure are described in detail in the Method section.

Accordingly, this study aimed to: (1) describe students' use and responses to an AI-Based Learning Assistant during TKJ learning; (2) examine changes in students' HOTS across repeated measurement occasions; and (3) determine whether the observed changes in HOTS across the measurement occasions were statistically significant and accompanied by a meaningful normalized learning gain. Given the one-group time-series design, the statistical hypothesis was formulated as follows: H0: there is no significant difference in students' mean HOTS scores across the measurement occasions; H1: at least one measurement occasion differs significantly from another. The findings are expected to contribute empirical evidence concerning the use of AI-based learning assistance in vocational education while recognizing the exploratory and single-group nature of the research design.

Method

This study used a quantitative approach with an experimental method. The design employed was pre-experimental in the form of a time series design, in which repeated measurements were taken on the research subjects both before and after the treatment. In this design, the subjects were first given an initial test (pretest) to establish a stable baseline condition, then given the treatment, and subsequently given a final test (posttest) to observe the changes that occurred continuously after the treatment. The quantitative approach was chosen because the study focused on measuring numerical data students' Higher Order Thinking Skills (HOTS) scores before and after the implementation of the AI-Based Learning Assistant that could be analyzed statistically, allowing the researcher to objectively measure the extent of the improvement in students' HOTS as an effect of the treatment (Sudianti et al., 2025).

Research Design and Procedure

This study employed a quantitative approach using a pre-experimental one-group time-series design. The design was selected to examine changes in students' Higher-Order Thinking Skills (HOTS) across repeated measurement occasions during the implementation of an AI-Based Learning Assistant. A single intact class participated in the intervention, and HOTS was measured repeatedly before, during, and after the intervention. The design therefore provides within-group evidence of change over time rather than a direct comparison between AI-assisted and conventional instruction.

The research design followed the sequence $O_1 \rightarrow X_1 \rightarrow O_2 \rightarrow X_2 \rightarrow O_3 \rightarrow \ldots \rightarrow X_n \rightarrow O_n$, where O represents HOTS measurement and X represents AI-assisted learning using ChatGPT. The repeated observations consisted of an initial pretest, periodic/intermediate tests, and a posttest. The repeated-measurement structure was intended to provide a more detailed description of the trajectory of students' HOTS than a single pretest-posttest comparison.

Because the study did not include a comparison group, the repeated measurements were not treated as a statistical control for maturation, history, or testing effects. Instead, the multiple measurement points were used to examine whether the observed pattern of HOTS change was consistent across the intervention period. Potential threats to internal validity, including maturation, history, repeated-testing/practice effects, and other time-related influences, were therefore considered when interpreting the findings. Consequently, the study's results are interpreted as evidence of significant within-group change associated with the intervention rather than definitive evidence of a causal effect attributable exclusively to ChatGPT.:

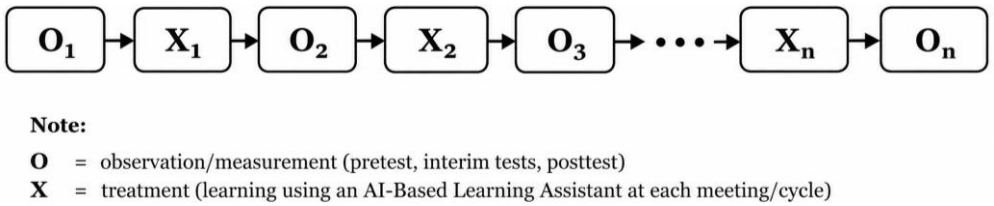

Figure 1. One-Group Time Series Research Design

The overall research procedure followed five main stages, from instrument preparation to the drawing of conclusions, as illustrated in Figure 2.

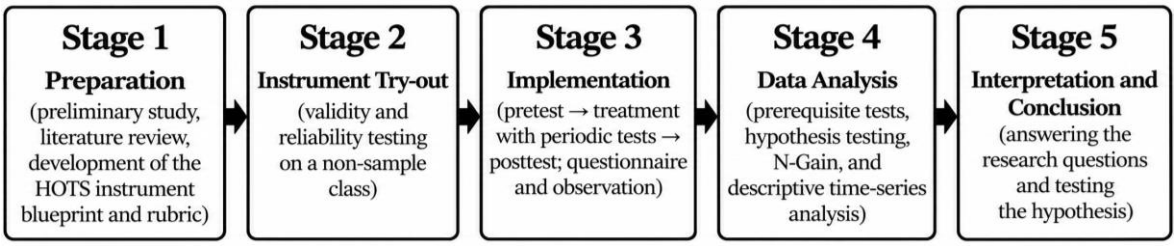

Figure 2. Flowchart of the Research Procedure One-Group Time Series Design

The overall research procedure followed five main stages: (a) the Preparation Stage, comprising a preliminary study, literature review, and the construction of the blueprint, items, and scoring rubric of the HOTS instrument based on the Learning Outcomes (CP) and Learning Objectives (TP) of the Cabled and Wireless Network Technology subject; (b) the Instrument Try-out Stage, in which the constructed instrument was piloted on a non-sample class (Grade XII) to determine its validity and reliability, with items that did not meet the requirements revised or discarded — the resulting coefficients are reported in full in the Results section (Tables 5–15), including content validity (Aiken's V), inter-rater agreement, item validity, item discrimination, and reliability (Cronbach's α) for all three instruments, so that the reader can verify instrument quality before interpreting the substantive findings; (c) the Implementation Stage, following the time series design, beginning with a pretest (O1), followed by treatment (X) in the form of learning using the AI-Based Learning Assistant delivered gradually and interspersed with periodic tests (O2, O3), and ending with a posttest (On), with the questionnaire and learning-activity observation conducted alongside the treatment; (d) the Data Analysis Stage, in which the pretest, periodic test, posttest, questionnaire, and observation data were analyzed through prerequisite tests, hypothesis testing (including a sphericity check and effect-size estimation, described below), the N-Gain test, and descriptive time-series analysis; and (e) the Interpretation and Conclusion Stage. Following this procedure, what was produced in this study was not a new AI application or system, but rather an experimental design together with a set of measurement instruments (a HOTS test, a questionnaire, and an observation guide) used to empirically examine changes in HOTS associated with the use of an already-available AI-Based Learning Assistant (ChatGPT) (Habib, Desky, Maksum, et al., 2026).

Research Site and Time

This research was conducted at SMK Muhammadiyah 1 Padang, with the research subjects being eleventh-grade students of the Computer and Network Engineering (TKJ) study program. The site was selected based on the relevance between the students' characteristics and the research focus, namely the implementation of an AI-Based Learning Assistant (ChatGPT) to measure its effectiveness in strengthening HOTS. The study was conducted during the current academic year, beginning with instrument development, followed by validity and reliability testing, and continuing with data collection through the pretest, the treatment (use of the AI-Based Learning Assistant in learning), and the posttest.

Population and Sample

The sampling frame consisted of Grade XI students enrolled in the Computer and Network Engineering (TKJ) study program at SMK Muhammadiyah 1 Padang during the 2026 academic year. The study involved 15 students from one intact TKJ class. Purposive sampling was used because the selected class met the predefined inclusion criteria: (1) students were enrolled in the Cabled and Wireless Network Technology course; (2) the class had access to the technological facilities required for AI-assisted learning; and (3) the class had not previously received systematic instruction using ChatGPT as an AI-Based Learning Assistant.

All 15 students in the selected class participated in the intervention and completed the repeated HOTS assessments.

Research Instrument

The instrument used to measure students' HOTS in Cabled and Wireless Network Technology learning referred to Anderson and Krathwohl's (2001) revised Bloom's taxonomy, covering the abilities of analysis (C4), evaluation (C5), and creation (C6). The instrument took the form of a problem-based essay test constructed from the Learning Outcomes (CP) and Learning Objectives (TP) of the subject, designed to measure students' ability to analyze problems, evaluate solutions, and design real-world solutions. The complete blueprint of the HOTS instrument is presented in Table 2.

Table 2. Blueprint of the HOTS Instrument

Element	Indicator of HOTS	Cognitive Level	Item Form
Cabled & Wireless Network Installation (Indoor/Outdoor)	Analyzing cable-type selection, cabling topology, supporting devices, and access-point placement for indoor/outdoor network installation	C4 (Analysis)	Contextual case-based essay (Pretest)
	Evaluating the feasibility of topology selection, wireless standards (802.11), and cable/wireless installation completeness according to installation SOP	C5 (Evaluation)	Contextual case-based essay (Pretest)
	Designing installation SOPs and layout/specifications for cable and access-point placement (indoor & outdoor) according to standards	C6 (Creation)	Contextual case-based essay (Posttest)
Cabled & Wireless Network Maintenance (Indoor/Outdoor)	Analyzing causes of connection disruption and wireless signal-quality degradation due to device age, distance, obstacles, or heavy device load	C4 (Analysis)	Contextual case-based essay (Pretest)
	Evaluating the feasibility of cable/access-point replacement or repair and the suitability of periodic maintenance schedules based on monitoring results	C5 (Evaluation)	Contextual case-based essay (Pretest)
	Designing preventive-maintenance SOPs, checklists, and escalation procedures for cabled and wireless network disruptions	C6 (Creation)	Contextual case-based essay (Posttest)
Cabled & Wireless Network Testing (Indoor/Outdoor)	Analyzing the needs of network testing and performance parameters (wiremap, cable length & continuity, throughput, signal strength, SNR)	C4 (Analysis)	Contextual case-based essay (Pretest)
	Evaluating cable and wireless test results against established standards/specifications as a basis for repair decisions	C5 (Evaluation)	Contextual case-based essay (Pretest)
	Designing a comprehensive testing procedure, test-result report, and post-operational periodic testing schedule for the network	C6 (Creation)	Contextual case-based essay (Posttest)

Each HOTS item was scored using a rubric ranging from 1 to 4, as shown in Table 3, based on four assessed aspects: (1) accuracy of problem analysis, (2) strength of argumentation, (3) accuracy of evaluation, and (4) creativity of the proposed solution.

Table 3. HOTS Scoring Rubric

Score	Criteria
4	Answer is very precise, analysis is in-depth, argumentation is strong, and the solution is creative
3	Answer is precise, analysis is fairly good, argumentation is logical
2	Answer is less precise, analysis is limited
1	Answer is inaccurate or does not demonstrate a thinking process

Instrument Try-out
Before being used in the main study, the HOTS essay-test instrument was piloted to ensure its quality in measuring higher-order thinking validly and consistently. Item validity was examined using the Pearson Product-Moment correlation between each item score and the total score:

$$rxy = [N\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{\{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]\}}$$

where rxy is the item-total correlation coefficient, X is the item score, Y is the total score, and N is the number of respondents; an item was declared valid if rcount > rtable. Instrument reliability was examined using Cronbach's Alpha:

$$\alpha = [k/(k-1)] \times [1 - (\Sigma Si^2 / St^2)]$$

where α is the reliability coefficient, k is the number of items, Si^2 is the variance of each item, and St^2 is the total variance; an instrument was declared reliable if $\alpha \geq 0.70$.

Data Collection Techniques
Data were collected through four complementary techniques following a time-series pattern. (1) Tests were used as the primary instrument to measure HOTS, comprising a pretest (administered before the treatment to establish a baseline), periodic/intermediate tests (administered during the learning process across several sessions to monitor gradual development), and a posttest (administered after the entire series of AI-assisted learning sessions to determine the final level of achievement). (2) A questionnaire was used to collect data on the level of use of and student responses to the AI-Based Learning Assistant, covering ease of use, engagement in learning, support for content comprehension, and perceived contribution to HOTS, using a five-point Likert scale as presented in Table 4. (3) Structured observation was conducted during each learning session to capture students' actual classroom activity that could not be fully captured through tests and questionnaires, focusing on students' activity in using AI, interaction in learning, and engagement in HOTS-based tasks. (4) Documentation was used to complement and strengthen the data obtained from the other techniques, including student test scores at each measurement stage, questionnaire recapitulations, learning-process notes, and evidence of AI-Based Learning Assistant use during learning activities.

Table 4. Five-Point Likert Scale

Score	Category
5	Strongly Agree (SA)
4	Agree (A)
3	Neutral (N)
2	Disagree (D)
1	Strongly Disagree (SD)

Data Analysis Technique
Quantitative data from the tests and the questionnaire were analyzed using several statistical techniques. First, a normality test was conducted using the Kolmogorov-Smirnov method with the Lilliefors significance correction (n = 15, α = 0.05) to determine whether the learning-outcome data at each measurement point were normally distributed; the data were declared normal if the computed D statistic was smaller than the D-table value. Second, because the design involved repeated measurement of the same subjects across three time points (pretest, periodic tests, and posttest), a repeated measures one-way ANOVA was used to determine whether a statistically significant mean difference existed across the measurement occasions, with H0 rejected if Sig. < 0.05 or Fcount > Ftable. Third, the N-Gain (Normalized Gain) test was used to determine the magnitude of improvement in students' HOTS following the treatment, calculated as:

$$g = (posttest - pretest) / (maximum\ score - pretest)$$

with the resulting g value categorized as high ($g \geq 0.70$), moderate ($0.30 \leq g < 0.70$), or low ($g < 0.30$). In addition, a descriptive time-series analysis was performed by computing the mean score and the percentage improvement at each measurement occasion and presenting the results in tabular and graphical form, so as to reveal whether students' HOTS increased, decreased, or remained stagnant across sessions. Questionnaire data were analyzed descriptively using the Likert scale, by computing each respondent's score, summing the total score, calculating the mean, and interpreting the result into a percentage category following Sugiyono (2016) and Riduwan (2015): 81–100% = Very Good, 61–80% = Good, 41–60% = Fair, 21–40% = Poor, and 0–20% = Very Poor.

Results and Discussions

Content Validity and Inter-Rater Reliability of the Instruments

Before being used for data collection, all three research instruments the HOTS test, the student-response questionnaire, and the observation sheet were reviewed for content validity by five expert validators. Each validator rated every item on a 1–5 scale, and the scores were processed using the Aiken's V formula, with V > 0.80 categorized as Very Valid, 0.40 – 0.80 as Valid, and V ≤ 0.40 as Less Valid. The results for each instrument are presented in Tables 5–7.

Table 5. Content Validity (Aiken's V) of the HOTS Test Instrument

Aspect Assessed	Number of Items	Mean Aiken's V	Category
Content/Material	10	0.860	Very Valid
Construction	10	0.860	Very Valid
Language	10	0.820	Very Valid
Pretest–Posttest Item Fit (Sessions 1–3)	60	0.849	Very Valid
Scoring Rubric Review	5	0.920	Very Valid
TOTAL	95	0.852	Very Valid

All aspects of the HOTS test instrument obtained a mean Aiken's V above 0.80 and were therefore categorized as Very Valid, with an overall mean of 0.852. Of the 95 items validated, 63 were categorized Very Valid and 32 Valid, while none were categorized Less Valid; the lowest Aiken's V value obtained was 0.70 and the highest reached 1.00, indicating that all HOTS items were appropriate for measuring students' higher-order thinking without requiring substantial revision.

Table 6. Content Validity (Aiken's V) of the Student-Response Questionnaire

Aspect Assessed	Number of Items	Mean Aiken's V	Category
Construction	6	0.867	Very Valid
Language	4	0.838	Very Valid
Content/Construct Fit	4	0.875	Very Valid
Item–Statement Fit Review	19	0.839	Very Valid
TOTAL	33	0.848	Very Valid

The mean Aiken's V for the four aspects of the questionnaire ranged from 0.838 to 0.875, with an overall mean of 0.848 (Very Valid). Of the 33 statements assessed, 20 were categorized Very Valid and 13 Valid, so that the entire questionnaire was declared appropriate for measuring student responses to learning.

Table 7. Content Validity (Aiken's V) of the Observation Sheet

Aspect Assessed	Number of Items	Mean Aiken's V	Category
Construction	4	0.812	Very Valid
Language	2	0.850	Very Valid
Content Fit	3	0.817	Very Valid
Observation-Indicator Fit Review	8	0.856	Very Valid
TOTAL	17	0.838	Very Valid

The mean Aiken's V for the observation sheet ranged from 0.812 to 0.856, with an overall mean of 0.838. Of the 17 indicators assessed, 10 were categorized Very Valid and 7 Valid. Taken together, all three instruments the HOTS test, the student-response questionnaire, and the observation sheet — were declared content-valid and appropriate for data collection. In addition to content validity, inter-rater reliability was examined using Cronbach's Alpha across the five validators' scores, with the results shown in Table 8.

Table 8. Inter-Rater Reliability (Cronbach's Alpha)

Instrument	Number of Items	Σ Variance per Validator	Total Item-Score Variance	Cronbach's Alpha	Category
HOTS Test	95	1.518	1.871	0.235	Less Reliable
Student-Response Questionnaire	33	1.366	1.843	0.324	Less Reliable
Observation Sheet	17	1.382	1.566	0.147	Less Reliable

The inter-rater Cronbach's Alpha values were low for all three instruments (0.235, 0.324, and 0.147, respectively), which, following Guilford's guideline, falls in the Less Reliable category. This low alpha must be interpreted cautiously: because the five validators' ratings on nearly all items were consistently clustered in the high range (4-5), the between-validator variance was very small (restriction of range), which mathematically depresses Cronbach's Alpha even though the validators' judgments were substantively consistent. To complement this information, a percent-agreement analysis was conducted, examining the proportion of items receiving identical scores or scores differing by no more than one point across the five validators, as shown in Table 9.

Table 9. Percent Agreement Among Validators

Instrument	Number of Items	% Identical Scores	% Difference ≤ 1 Point	Agreement Interpretation
HOTS Test	95	7.4	87.4	High
Student-Response Questionnaire	33	6.1	87.9	High
Observation Sheet	17	29.4	100.0	Very High

The percentage of items with a between-validator score difference of no more than one point reached 87.4% for the HOTS test, 87.9% for the student-response questionnaire, and 100% for the observation sheet, corresponding to High to Very High agreement categories. This finding reinforces the conclusion that, despite the low Cronbach's Alpha caused by the limited variance among validator scores, the five validators practically demonstrated a high level of agreement across all instruments. Considering the content-validity results together with the inter-rater agreement analysis, the HOTS test, the student-response questionnaire, and the observation sheet were declared valid and appropriate for use in this study.

Empirical Try-out of the HOTS Test (Grade XII Try-out Class)

Before being used with the main research subjects (Grade XI), the HOTS test instrument was first piloted empirically with 15 Grade XII students to statistically examine item difficulty, item validity, discrimination power, and reliability. The piloted instrument consisted of 60 HOTS items distributed across three learning topics (Topic 1, 2, and 3), each comprising 20 items divided into Part A (10 items) and Part B (10 items), scored on a 0–4 rubric scale, yielding an ideal maximum score of 240 per student. The descriptive results of the try-out are presented in Table 10.

Table 10. Descriptive Summary of the Grade XII HOTS Try-out Scores

N	Minimum Score	Maximum Score	Mean Score	Mean Percentage (%)
15	71	155	124.2	51.75

Of the 15 students, the lowest total score obtained was 71 and the highest was 155, with a mean total score of 124.2 out of an ideal maximum of 240 (51.75%). Based on total scores, students were subsequently grouped into upper, middle, and lower groups, with the top 27% (4 students) and the bottom 27% (4 students) used for the item-discrimination analysis. The difficulty index (IK) of each item was calculated as the mean item score divided by the maximum item score (4), and categorized as easy (IK ≥ 0.70), moderate (0.30–0.69), or difficult (IK < 0.30); the results are shown in Table 11.

Table 11. Summary of Item Difficulty Level

Difficulty Category	Number of Items	Percentage (%)
Easy (IK ≥ 0.70)	0	0.0
Moderate (0.30–0.69)	60	100.0
Difficult (IK < 0.30)	0	0.0

All 60 piloted items fell into the moderate difficulty category, with none categorized as easy or difficult, indicating that the item set was fairly homogeneous in terms of difficulty and did not yet reflect a varied difficulty spread. Item validity was analyzed using the Product-Moment correlation between each item score and the total score, compared against rtable at n = 15 and α = 5% (0.514); the results are shown in Table 12.

Table 12. Summary of Item Validity

Validity Category	Number of Items	Percentage (%)
Valid	37	61.7
Invalid	23	38.3

Of the 60 piloted items, 37 were declared valid and 23 were declared invalid. The discrimination power (DP) of each item was then calculated as the difference between the upper- and lower-group mean scores divided by the maximum item score, categorized as good (DP ≥ 0.40), fair (0.20–0.39), poor (0.00–0.19), or very poor (negative); the results are shown in Table 13.

Table 13. Summary of Item Discrimination Power

Discrimination Category	Number of Items	Percentage (%)
Good (DP ≥ 0.40)	5	8.3
Fair (0.20–0.39)	32	53.3
Poor (0.00–0.19)	22	36.7
Very Poor (negative)	1	1.7

Most items (32 items, 53.3%) fell into the fair discrimination category, while only 5 items (8.3%) were categorized as good; 22 items (36.7%) were categorized poor and 1 item (1.7%) very poor — meaning that item was actually answered correctly more often by the lower group than the upper group. Combining the validity and discrimination results yielded a follow-up recommendation for each item, summarized in Table 14.

Table 14. Summary of Item Follow-up Recommendations

Recommendation	Number of Items	Percentage (%)
Retained for use	37	61.7
Discarded / replaced	23	38.3

Of the 60 piloted items, 37 (61.7%) were declared fit for use, while 23 (38.3%) were recommended for removal or replacement because they did not meet the validity and/or discrimination criteria; these findings became the main consideration in revising the HOTS instrument prior to data collection with the Grade XI subjects. Instrument reliability was also examined using Cronbach's Alpha, both for the full 60-item set and for each topic sub-group, as shown in Table 15.

Table 15. Reliability (Cronbach's Alpha) of the Try-out Instrument

Item Group	Number of Items	Alpha Value	Interpretation
Overall	60	0.934	Very High
Topic 1	20	0.905	Very High
Topic 2	20	0.829	High (Reliable)
Topic 3	20	0.883	High (Reliable)

The Cronbach's Alpha for the full 60-item set was 0.934 (Very High); per topic, Topic 1 obtained 0.905 (Very High), Topic 2 obtained 0.829 (High/Reliable), and Topic 3 obtained 0.883 (High/Reliable). The HOTS test instrument was therefore declared reliable, both overall and within each topic group, even though not every item met the validity and discrimination criteria, underscoring the importance of the item-revision process before the instrument was used with the main research subjects.

Descriptive Statistics of Student Learning Outcomes

Data on student learning outcomes were obtained across three learning sessions (Session 1, Session 2, and Session 3) involving 15 Grade XI students. After the item-validity analysis, 23 of the 60 items were declared invalid and excluded from subsequent scoring, resulting in an ideal maximum score of 52 for Session 1 and Session 3 and 44 for Session 2. The descriptive statistics of the three sessions are presented in Table 16.

Table 16. Descriptive Statistics of Learning Outcomes

	N	Minimum	Maximum	Mean	Std. Deviation
Session 1 Score	15	55.77	76.92	66.41	5.622
Session 2 Score	15	59.09	81.82	72.88	5.533
Session 3 Score	15	63.46	86.54	77.56	6.961

Student learning outcomes increased across sessions: the mean score rose from 66.41 in Session 1 (min 55.77, max 76.92) to 72.88 in Session 2 (min 59.09, max 81.82), and reached 77.56 in Session 3 (min 63.46, max 86.54) — an increase of 6.47 points from Session 1 to Session 2, a further 4.68-point increase to Session 3, and an overall 11.15-point increase from Session 1 to Session 3. The standard deviation was 5.622 in

Session 1, decreased slightly to 5.533 in Session 2, and increased to 6.961 in Session 3, indicating that the spread of scores remained relatively controlled, although it widened somewhat by Session 3.

Test of Analysis Prerequisite (Normality)

Prior to hypothesis testing, a normality test was conducted using the Kolmogorov-Smirnov method with the Lilliefors correction (n = 15, α = 0.05) to determine whether the learning-outcome data at each session were normally distributed. The results are shown in Table 17.

Table 17. Normality Test Results

	Statistic (D)	df	D-table (α = 0.05)	Remark
Session 1 Score	0.133	15	0.220	Normal
Session 2 Score	0.084	15	0.220	Normal
Session 3 Score	0.137	15	0.220	Normal

The computed D statistics for Session 1 (0.133), Session 2 (0.084), and Session 3 (0.137) were all smaller than the D-table value of 0.220, so H0 was accepted and the data at all three sessions were declared normally distributed. The data therefore satisfied the normality assumption, and the analysis proceeded using parametric statistics, namely a repeated measures one-way ANOVA.

Hypothesis Testing

Because the data originated from the same subjects measured at three time points, a repeated measures one-way ANOVA was used to test the following hypotheses: H0 there is no significant mean difference in learning outcomes among Session 1, Session 2, and Session 3; H1 a significant mean difference exists in at least two of the three sessions. H0 was rejected if Sig. < 0.05 or Fcount > Ftable. The descriptive statistics are shown in Table 18 and the ANOVA results in Table 19.

Table 18. Descriptive Statistics of Learning Outcomes by Session

Session	Mean	Std. Deviation	N
Session 1	66.41	5.622	15
Session 2	72.88	5.533	15
Session 3	77.56	6.961	15

Table 19. Repeated Measures ANOVA (Within-Subjects Effects)

Source of Variance	Sum of Squares	df	Mean Square	F	Sig.
Between Sessions (Treatment)	940.430	2	470.215	54.993	0.000
Between Subjects (Students)	1310.904	14	93.636	–	–
Error	239.413	28	8.550	–	–
Total	2490.747	44	–	–	–

Ftable (df1 = 2, df2 = 28, α = 0.05) = 3.340. Criterion: H0 is rejected if Sig. < 0.05.

The repeated measures one-way ANOVA yielded Fcount = 54.993 with p < 0.001. Because Fcount exceeded Ftable (3.340) and the significance value was below 0.05, H0 was rejected, indicating a statistically significant mean difference in learning outcomes among Session 1, Session 2, and Session 3, consistent with the increasing pattern of mean scores from 66.41 to 72.88 to 77.56.

N-Gain Analysis

Following the ANOVA, the N-Gain test was used to determine the magnitude of improvement in learning outcomes, categorized as high (g ≥ 0.70), moderate (0.30 ≤ g < 0.70), or low (g < 0.30). The summary of N-Gain values for each session comparison is presented in Table 20.

Table 20. Summary of N-Gain Between Sessions

Comparison	Mean (Before)	Mean (After)	N-Gain (g)	Category
Session 1 → Session 2	66.41	72.88	0.191	Low
Session 2 → Session 3	72.88	77.56	0.186	Low
Session 1 → Session 3 (Overall)	66.41	77.56	0.335	Moderate

The improvement from Session 1 to Session 2 (g = 0.191) and from Session 2 to Session 3 (g = 0.186) each fell into the Low category; however, the overall improvement from Session 1 to Session 3 (g = 0.335) fell into the Moderate category. This indicates that, although the incremental gain at each individual stage was modest, the cumulative improvement in students' understanding across the full learning process was

meaningful and reached the moderate category. The detailed, student-level N-Gain values for the Session 1 → Session 2 and Session 1 → Session 3 comparisons are presented in Tables 21 and 22.

Table 21. Student-Level N-Gain, Session 1 → Session 2

Initial Score	Final Score	N-Gain (g)	Category
61.54	68.18	0.173	Low
69.23	70.45	0.040	Low
65.38	70.45	0.146	Low
63.46	70.45	0.191	Low
65.38	75.00	0.278	Low
71.15	75.00	0.133	Low
61.54	68.18	0.173	Low
59.62	72.73	0.325	Moderate
71.15	75.00	0.133	Low
76.92	81.82	0.212	Low
63.46	77.27	0.378	Moderate
55.77	59.09	0.075	Low
69.23	77.27	0.261	Low
71.15	72.73	0.055	Low
71.15	79.55	0.291	Low

Most students fell into the low category for the Session 1→Session 2 comparison; only two students, Gusti Clara (g = 0.325) and Mela Mardiati (g = 0.378), reached the moderate category, while the remaining 13 students fell into the low category, indicating that improvement at this early stage of learning was not yet evenly distributed among students.

Table 22. Student-Level N-Gain, Session 1 → Session 3 (Overall)

Initial Score	Final Score	N-Gain (g)	Category
61.54	71.15	0.250	Low
69.23	73.08	0.125	Low
65.38	73.08	0.222	Low
63.46	71.15	0.210	Low
65.38	78.85	0.389	Moderate
71.15	84.62	0.467	Moderate
61.54	75.00	0.350	Moderate
59.62	76.92	0.428	Moderate
71.15	75.00	0.133	Low
76.92	86.54	0.417	Moderate
63.46	86.54	0.632	Moderate
55.77	63.46	0.174	Low
69.23	84.62	0.500	Moderate
71.15	76.92	0.200	Low
71.15	86.54	0.533	Moderate

For the overall Session 1→Session 3 comparison, 8 of the 15 students reached the moderate category, while 7 remained in the low category; the highest N-Gain was achieved by Mela Mardiati (g = 0.632, moderate). This result shows that when improvement is observed from the beginning to the end of the learning process, more students reach the moderate category compared with the stage-by-stage comparisons, confirming the pattern observed in the mean-score analysis of a gradual improvement in learning outcomes from Session 1 through Session 3.

Observation of Student Activity and Interaction

In addition to the learning-outcome and questionnaire data, observation data were collected to directly capture student activity and interaction during learning. Observations were conducted by the researcher at every session using the content-validated observation sheet (Table 7), rated on a scale of 1 to 4 (1 = Not Observed, 2 = Slightly Observed, 3 = Observed, 4 = Strongly Observed), comprising 8 behavioral indicators grouped into three aspects: (1) AI-Use Activity, (2) Learning Interaction, and (3) Engagement in HOTS Tasks. The results across Session 1, 2, and 3 are presented in Table 23.

Table 23. Observation of Student Learning Activity and Interaction

Aspect	Indicator Observed	S1	S2	S3
AI-Use Activity	Students independently formulate prompts appropriate to the given case	1	3	3
	Students use AI output as discussion material rather than a final answer	3	3	4
	Students verify/critique the AI's answer	2	3	3
Learning Interaction	Students actively discuss with group members	3	3	3
	Students ask the teacher when they reach an impasse	3	4	4
Engagement in HOTS Tasks	Students actively engage in the problem-analysis stage	2	3	3
	Students actively engage in the solution-evaluation stage	2	3	3
	Students actively engage in the solution-design stage	3	3	3

Table 23 shows a rising observation score on nearly every indicator from Session 1 to Session 3, with no indicator declining. To provide a more concise overview of the trend on each aspect, the mean and achievement percentage per aspect are summarized in Table 24.

Table 24. Recapitulation of Mean Observation Scores by Aspect and Session

Aspect	Number of Indicators	Session 1	Session 2	Session 3
AI-Use Activity	3	2.00 (50.0%)	3.00 (75.0%)	3.33 (83.3%)
Learning Interaction	2	3.00 (75.0%)	3.50 (87.5%)	3.50 (87.5%)
Engagement in HOTS Tasks	3	2.33 (58.3%)	3.00 (75.0%)	3.00 (75.0%)
Overall (Mean)	8	2.38 (59.4%)	3.13 (78.1%)	3.25 (81.3%)

On the AI-Use Activity aspect, the mean observation score rose from 2.00 (50.0%) in Session 1 to 3.00 (75.0%) in Session 2, and reached 3.33 (83.3%) in Session 3; the most notable improvement occurred on the independent prompt-formulation indicator, which began at score 1 (Not Observed) as most students still relied on teacher direction, then rose to score 3 (Observed) in Session 2 and remained there in Session 3, reflecting growing student autonomy in using the AI-Based Learning Assistant. The verification/critique indicator likewise rose from score 2 to 3, suggesting that students increasingly evaluated the accuracy and relevance of AI output rather than accepting it passively (Mardiana et al., 2026).

On the Learning Interaction aspect, the score rose from 3.00 (75.0%) in Session 1 to 3.50 (87.5%) in Sessions 2 and 3, with group discussion and teacher consultation remaining consistently in the Observed to Strongly Observed range throughout, indicating that the AI-Based Learning Assistant did not replace social interaction in learning but rather operated alongside it students continued to discuss actively with peers and consult the teacher when facing difficulty, so that AI functioned more as a thinking aid than a substitute for direct instructional interaction (Masriani et al., 2026).

On the Engagement in HOTS Tasks aspect, the mean score rose from 2.33 (58.3%) in Session 1 to 3.00 (75.0%) in Session 2 and remained at 3.00 (75.0%) in Session 3. The three indicators on this aspect engagement in problem analysis, solution evaluation, and solution design correspond respectively to the analyzing, evaluating, and creating domains of Bloom's taxonomy, and the observed increase in engagement across these stages reinforces the HOTS test findings, so that the observation and learning-outcome data mutually corroborate one another. Overall, the mean observation score rose from 2.38 (59.4%) in Session 1 to 3.13 (78.1%) in Session 2 and reached 3.25 (81.3%) in Session 3, demonstrating a consistent upward trend in student activity and engagement across the learning process. It should be noted, however, that the observations in this study were conducted by a single observer at each session without inter-rater reliability testing, which leaves some possibility of scoring subjectivity that warrants attention in future research (Risna et al., 2026).

Student Questionnaire Responses

Beyond the learning-outcome data, student responses to the use of the AI-Based Learning Assistant were gathered through a 19-item, five-point Likert-scale questionnaire divided into four parts: Part A (items 1–5), Part B (items 6–10), Part C (items 11–14), and Part D (items 15–19), completed by all 15 research subjects after the full series of learning sessions. The percentage-score interpretation followed Sugiyono (2016) and

Riduwan (2015): 81–100% = Very Good, 61–80% = Good, 41–60% = Fair, 21–40% = Poor, and 0–20% = Very Poor. The recapitulation by questionnaire part is presented in Table 25.

Table 25. Recapitulation of Student Responses by Questionnaire Part

Part	Item Range	Mean Score	Mean Percentage	Category
Part A	1–5	60.80	81.07%	Very Good
Part B	6–10	60.40	80.53%	Good
Part C	11–14	59.25	79.00%	Good
Part D	15–19	59.40	79.20%	Good
Overall	1–19	–	80.00%	Good

Overall, student responses to the use of the AI-Based Learning Assistant fell into the Good category with a mean percentage of 80.00%. Part A, concerning the ease of use and usefulness of the AI-Based Learning Assistant, obtained the highest percentage (81.07%, Very Good), while Parts B, C, and D obtained 80.53%, 79.00%, and 79.20% respectively (Good). No part fell into the Fair, Poor, or Very Poor category, indicating that students generally responded positively to the AI-Based Learning Assistant across all measured aspects. The item-level detail is presented in Table 26.

Table 26. Recapitulation of Student Responses by Item

Part	Total Score	Ideal Max. Score	Percentage	Category
A	62.0	75.0	82.67%	Very Good
A	59.0	75.0	78.67%	Good
A	62.0	75.0	82.67%	Very Good
A	60.0	75.0	80.00%	Good
A	61.0	75.0	81.33%	Very Good
B	57.0	75.0	76.00%	Good
B	59.0	75.0	78.67%	Good
B	64.0	75.0	85.33%	Very Good
B	62.0	75.0	82.67%	Very Good
B	60.0	75.0	80.00%	Good
C	56.0	75.0	74.67%	Good
C	60.0	75.0	80.00%	Good
C	61.0	75.0	81.33%	Very Good
C	60.0	75.0	80.00%	Good
D	59.0	75.0	78.67%	Good
D	57.0	75.0	76.00%	Good
D	60.0	75.0	80.00%	Good
D	60.0	75.0	80.00%	Good
D	61.0	75.0	81.33%	Very Good

Of the 19 questionnaire items, 7 fell into the Very Good category (items 1, 3, 5, 8, 9, 13, and 19) and 12 into the Good category, with none in the Fair, Poor, or Very Poor category. The highest percentage was obtained by item 8 in Part B (85.33%), and the lowest by item 11 in Part C (74.67%), which nonetheless remained in the Good category. Responses were further analyzed per respondent, as shown in Table 27.

Table 27. Recapitulation of Student Responses by Respondent

Total Score	Ideal Max. Score	Percentage	Category
69	95	72.63%	Good
76	95	80.00%	Good
69	95	72.63%	Good
70	95	73.68%	Good
72	95	75.79%	Good
84	95	88.42%	Very Good
86	95	90.53%	Very Good
78	95	82.11%	Very Good
73	95	76.84%	Good
76	95	80.00%	Good
76	95	80.00%	Good

Total Score	Ideal Max. Score	Percentage	Category
74	95	77.89%	Good
79	95	83.16%	Very Good
78	95	82.11%	Very Good
80	95	84.21%	Very Good

Of the 15 students, 6 gave responses in the Very Good category Radit May Hendra (88.42%), Zaidina Berlian (90.53%), Gusti Clara (82.11%), Nela Julia (83.16%), Rio Firmansyah (82.11%), and Khaira Nur Fadhilah (84.21%) while the remaining 9 students responded in the Good category, ranging from 72.63% to 80.00%. No student responded in the Fair, Poor, or Very Poor category. The highest individual response was recorded for Zaidina Berlian (90.53%), and the lowest for Cancan and Rahul Putra (72.63% each), both still within the Good category. Overall, the questionnaire recapitulation indicates that all students responded positively to the use of the AI-Based Learning Assistant, with acceptance ranging from Good to Very Good.

Summary of Statistical Testing

Table 28. Summary of Hypothesis-Testing Results

Aspect Tested	Result	Remark
Normality Test (Lilliefors)	D 0.220< for all three sessions	Data normally distributed
Repeated Measures One-Way ANOVA	F = 54.993; Sig. = 0.000 (< 0.05)	H0 rejected; significant difference
N-Gain Test (Overall)	g = 0.335	Moderate category

The findings above are discussed in relation to the theoretical framework and prior relevant studies, organized around the three research questions: (1) the level of use of the AI-Based Learning Assistant in Computer and Network Engineering (TKJ) learning; (2) students' HOTS before and after using the AI-Based Learning Assistant; and (3) the effectiveness of the AI-Based Learning Assistant in strengthening the HOTS of students at SMK Muhammadiyah 1 Padang.

Level of Use of the AI-Based Learning Assistant. Based on the questionnaire recapitulation in Table 25, students' overall response to the AI-Based Learning Assistant fell into the Good category (80.00%), with Part A concerning ease of use and usefulness obtaining the highest percentage (81.07%, Very Good) and Parts B, C, and D obtaining 80.53%, 79.00%, and 79.20% respectively (Good). No part of the questionnaire fell below the Good category, indicating that students generally accepted and made good use of the AI-Based Learning Assistant throughout the learning process. This finding is reinforced by the item- and respondent-level analyses in Tables 26 and 27, which show that none of the 19 items and none of the 15 respondents fell below the Good category even the lowest item (74.67%) and the lowest individual respondents (72.63%) remained within the Good range suggesting a fairly even level of acceptance across the class, without any subgroup of students displaying meaningful rejection of, or difficulty adapting to, the technology.

This Good level of use is consistent with the view of (Sari et al., 2026) and (Kartini et al., 2026), who argue that AI tools such as ChatGPT enable more individualized learning because each student receives challenges matched to their cognitive level, making the learning experience more focused and productive. The finding is also consistent with (Ningsih et al., 2025), who found ChatGPT effective in supporting learning among vocational students. Nevertheless, this high level of acceptance still needs to be balanced with teacher supervision and guidance, given (Pebrianti et al., 2026) caution that undirected AI use has the potential to reduce students' critical-thinking capacity through over-reliance on the technology.

Students' HOTS Before and After Using the AI-Based Learning Assistant. Before addressing the second research question, it must be emphasized that the HOTS data analyzed in this study passed through a rigorous instrument-quality process. The Aiken's V content-validity results in Table 5 show that every aspect of the HOTS test instrument reached the Very Valid category (mean 0.852), reinforced by the high inter-validator agreement in Table 9. The subsequent empirical try-out with Grade XII students (Tables 11–15) showed very high reliability ($\alpha = 0.934$), although of the 60 piloted items only 37 (61.7%) met the validity and discrimination criteria, while 23 (38.3%) were recommended for removal or revision. This item-selection process was essential to ensure that the scores analyzed across Sessions 1–3 genuinely reflected students' HOTS rather than being an artifact of weak items (Virnindo et al., 2026).

Based on the descriptive results in Tables 16 and 18, students' HOTS improved consistently across sessions, from a mean of 66.41 in Session 1 to 72.88 in Session 2 and 77.56 in Session 3. This gradual, stage-by-stage pattern of improvement indicates that students' ability to analyze, evaluate, and create solutions to

Computer and Network Engineering problems did not develop instantly, but rather progressed alongside the increasing intensity of student interaction with the AI-Based Learning Assistant across problem-based tasks in each session.

The N-Gain results in Table 20 reinforce this pattern: the Session 1→Session 2 (g = 0.191) and Session 2→Session 3 (g = 0.186) comparisons each fell into the Low category, while the overall Session 1→Session 3 improvement (g = 0.335) fell into the Moderate category. This pattern is consistent with the scaffolding concept within Vygotsky's socio-constructivist learning theory, as cited by Ardianti et al. in (Zhang et al., 2026), whereby higher-order thinking ability develops gradually through sustained guidance in this case, guidance provided by the AI-Based Learning Assistant. Nevertheless, the magnitude of improvement observed in this study was more moderate than that reported by (Labadze et al., 2024), who obtained an N-Gain of 0.83 (high category) from ChatGPT-assisted Problem-Based Learning; this difference is presumably related to the shorter intervention duration in the present study (only three sessions), which may not have given students sufficient time to fully adapt higher-order thinking patterns through AI interaction.

Effectiveness of the AI-Based Learning Assistant in Strengthening HOTS. The third research question was addressed through hypothesis testing using a repeated measures one-way ANOVA, conducted after the data were confirmed to be normally distributed at all three sessions (Table 17). The results in Table 19 show Fcount = 54.993 with $p < 0.001$, far below 0.05, so that H0 was rejected and H1 accepted indicating a statistically significant difference in mean HOTS scores among Session 1, Session 2, and Session 3, and demonstrating that the use of the AI-Based Learning Assistant was statistically effective in strengthening the HOTS of TKJ students at SMK Muhammadiyah 1 Padang.

This finding empirically supports the conceptual framework of the study, namely that student interaction with the AI-Based Learning Assistant analyzing information, comparing solutions, and evaluating outcomes indirectly trains students' critical and creative-thinking abilities (Wu et al., 2026). It is also consistent with (Ermawalis et al., 2025), who found a significant, moderate correlation (r = 0.641) between AI use and critical thinking among students of SMK Negeri 33 Jakarta, and with (Deviana et al., 2025), who found that ChatGPT as a virtual tutor improved students' HOTS, particularly in analysis and problem solving.

Nevertheless, this statistical significance must be interpreted proportionally. The overall N-Gain value in the Moderate category (0.335) indicates that the improvement in HOTS, while real, has not yet reached the high category. In addition, the finding that 38.3% of the initial HOTS items did not meet the validity and discrimination criteria (Tables 12–14) suggests that the effectiveness of an AI-Based Learning Assistant cannot be separated from the quality of the accompanying instrument and instructional design. This is consistent with (Fiona et al., 2025) observation that AI can help students answer HOTS items well but still requires validation and teacher/researcher understanding for its use to be well-targeted, as well as (Irfan et al., 2025) warning that AI has the potential to reduce critical-thinking ability if used without adequate direction.

Taken together, the findings on the three research questions complement one another and form a coherent conclusion: students accepted and used the AI-Based Learning Assistant well (Tables 25–27); students' HOTS improved gradually and consistently from the first to the third session (Tables 16, 18, 20–22); and this improvement was statistically significant (Table 19). Accordingly, the alternative hypothesis (H1) that an effective and significant improvement in students' HOTS occurred following the use of the AI-Based Learning Assistant in Computer and Network Engineering learning is accepted, while the null hypothesis (H0) is rejected. Even so, the limited number of subjects (15 students), the relatively short intervention duration (three sessions), and the need for further refinement of the HOTS test instrument remain important considerations, discussed further as limitations of the study below (Tananda et al., 2026).

Limitations of the Study

Several limitations should be acknowledged when interpreting these findings. First, the study involved only 15 students from a single intact class at one vocational school without a comparison (control) group, so that threats to internal validity including maturation, history, and testing effects across the repeated measurement occasions cannot be entirely ruled out. Second, the absence of random assignment and the single-school implementation restrict the external validity and generalizability of the findings to other vocational schools, study programs, and student populations. Third, the intervention was limited to three learning sessions, which may not have been sufficient for students to fully internalize higher-order thinking habits through AI-assisted interaction, as reflected in the still-moderate overall N-Gain. Fourth, the observation data were collected by a single observer per session without inter-rater reliability testing, leaving room for subjectivity in scoring. Future studies are therefore recommended to employ larger, multi-school samples with a quasi-

experimental or randomized comparison-group design, extend the intervention duration, and strengthen inter-rater reliability procedures for observational data, so as to establish more robust evidence of the causal effectiveness and broader generalizability of AI-Based Learning Assistants in vocational HOTS development.

Conclusions

Based on the results of the research and discussion, it can be concluded that: (1) the level of use of the AI-Based Learning Assistant in Computer and Network Engineering (TKJ) learning at SMK Muhammadiyah 1 Padang fell into the Good category, with a mean student-response percentage of 80.00%, and every aspect measured ease of use, usefulness, and support for the learning process obtained a Good to Very Good category, with no part, item, or respondent falling into the Fair, Poor, or Very Poor category; (2) students' Higher Order Thinking Skills (HOTS) improved consistently from before to after the use of the AI-Based Learning Assistant, from a mean of 66.41 in Session 1 to 72.88 in Session 2 and 77.56 in Session 3, with an overall N-Gain of 0.335 (moderate category); and (3) hypothesis testing using a repeated measures one-way ANOVA yielded Fcount = 54.993 with $p < 0.001$, so that H0 was rejected and H1 accepted, confirming that the use of the AI-Based Learning Assistant was effective and statistically significant in strengthening the HOTS of students at SMK Muhammadiyah 1 Padang in Computer and Network Engineering learning. Thus, an AI-Based Learning Assistant such as ChatGPT is worthy of recommendation as a supplementary instructional medium for vocational teachers and students, particularly to support independent, problem-based, and higher-order thinking-oriented learning, while remaining mindful of the need for adequate teacher guidance and continued refinement of measurement instruments.

Acknowledgments

Thank you to the Dean of FT UNP, Chair of the Master of Education and Kejuraun Study Program, lecturers and students of Electrical Engineering Education at Padang State University and all parties so that this research can be carried out well.

References

Aksara, W. D. A. N., & Gangga, A. (2026). *CIPP-Based Evaluation of the Learning Process in Vocational High Schools : Student and Teacher Perspectives*. *6*(1), 76–90.

Arwizet, K., Purwanto, W., Dewi, I. P., Habib, A., & Desky, A. (2026). *Development of a Web-Based CBT Worksheet for Routine Motorcycle Maintenance*. *12*(6), 223–230. https://doi.org/10.29303/jppipa.v12i6.15243

Aulia, L. N., Susilo, S., & Subali, B. (2019). Upaya peningkatan kemandirian belajar siswa dengan model problem-based learning berbantuan media Edmodo. *Jurnal Inovasi Pendidikan IPA*, *5*(1), 69–78. https://doi.org/10.21831/jipi.v5i1.18707

Chiu, T. K. F., Xia, Q., Zhou, X., Sing, C., & Cheng, M. (2023). Computers and Education : Artificial Intelligence Systematic literature review on opportunities , challenges , and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*(September 2022), 100118. https://doi.org/10.1016/j.caeai.2022.100118

Dedy, R., Budiman, A., Surjono, H. D., Firdaus, M., Kurniati, T., Feladi, V., Oktarika, D., Hakiki, M., Sabir, A., Wiyoko, T., Kadir, A., Hamid, M. A., & Fadli, R. (2025). *Effectiveness of AI-Driven Assessments in Enhancing Learning Evaluation through Predictive Technology in Vocational Secondary School*. *15*(7), 1410–1417. https://doi.org/10.18178/ijiet.2025.15.7.2342

Desky, A. H. A., Ganefri, G., Yulastri, A., Ta'ali, T., Effendi, H., & Fiandra, Y. A. (2025). Entrepreneurship-Based School Management Model to Increase Entrepreneurial Interest of Vocational Education Students at SMK Negeri 5 Padang. *Al Qalam: Jurnal Ilmiah Keagamaan Dan Kemasyarakatan*, *19*(1), 297. https://doi.org/10.35931/aq.v19i1.4306

Deviana, B., Zen, Z., Rayendra, J, F. Y., Mirshad, E., & Desky, A. H. A. (2025). Development of E-modules Based on Quantum Learning in Projek IPAS for Class X Vocational High School. *Jurnal Penelitian Pendidikan IPA*, *11*(9), 583–591. https://doi.org/10.29303/jppipa.v11i9.12179

Dodo, M., In, A., Zukhrufurrohmah, Z., & Ernawati, S. (2023). *Implementasi artificial intellegence ChatGPT melalui pembelajaran problem based learning dengan pendekatan TARL dalam meningkatkan kemampuan berpikir kritis siswa*. *0066*, 97–107.

Ermawalis, J, F. Y., Darmansyah, Rayendra, Mirshad, E., & Desky, A. H. A. (2025). Media in Improving

Teachers Professional Competence at SMAN 5 Batam. *Jurnal Penelitian Pendidikan IPA*, *11*(9), 592–598. https://doi.org/10.29303/jppipa.v11i9.12180

Fiona, Darmansyah, Zen, Z., J, F. Y., Mirshad, E., & Desky, A. H. A. (2025). The Effect of Applying the Archerys ID Model Instructional Design and Learning Motivation on Physics Learning Outcomes at SMAN 4 Batam. *Jurnal Penelitian Pendidikan IPA*, *11*(9), 609–619. https://doi.org/10.29303/jppipa.v11i9.12181

Fuad, M., & Kasman, R. A. (2026). *EFEKTIVITAS IMPLEMENTASI CHATGPT TERINTEGRASI STEM KEMAMPUAN BERPIKIR KRITIS SISWA pembelajaran Biologi . Proses pembelajaran di sekolah masih didominasi oleh metode konvensional yang pengetahuannya sendiri . Akibatnya , kesempatan peserta didik untuk melakukan investigasi , mengajukan berbagai teknologi pembelajaran telah tersedia [15]. 5*(1), 20–28.

Habib, A., Desky, A., Anwar, M., Simatupang, W., & Adri, M. (2026). *Development and Evaluation of an AI-Integrated Project-Based Learning Model for Technical and Vocational Education. 7*(2), 319–326.

Habib, A., Desky, A., Effendi, H., & Maksum, H. (2025). *Development of Project Based Learning E-Modules on Android Based Embedded System SubjectsHabib, A., Desky, A., Effendi, H., & Maksum, H. (2025). Development of Project Based Learning E-Modules on Android Based Embedded System Subjects. 11(10), 362–371. ht. 11*(10), 362–371. https://doi.org/10.29303/jppipa.v11i10.12358

Habib, A., Desky, A., Jalinus, N., & Ananda, Y. F. (2026). *Synergizing Technological Innovation , Quality Management , and Inclusivity in Transforming Industry Responsive Technical and Vocational Education and Training TVET. 7*(1), 40–46.

Habib, A., Desky, A., Maksum, H., Mukhaiyar, R., & Purwanto, W. (2026). *Development of Digital Learning Materials Based on Project- Based Learning Using CorelDRAW and Blender to Enhance Graphic Design and 3D Animation Skills. 12*(7), 823–835. https://doi.org/10.29303/jppipa.v12i7.15673

Habib, A., Desky, A., & Yulastri, A. (2026). *Validity of Animated Video-Based Learning Media Development Using Adobe After Effect Application in Grade IV Elementary School. 12*(2), 507–514. https://doi.org/10.29303/jppipa.v12i2.10424

Irfan, D., Wahyuni, T. S., Hadi, A., & Arrasyidi, A. H. (2025). *The Effect of Gamification in Computer Science Learning on Student Motivation and Learning Outcomes at SMP Negeri 27 Padang. 11*(9), 957–966. https://doi.org/10.29303/jppipa.v11i9.12543

Kartini, R., Irfan, D., Giatman, M., Wahyuni, T. S., Habib, A., & Desky, A. (2026). *The Effect of Literacy and Numeracy Skills and Field Work Experience on Employment Readiness Through Self-Efficacy Among Students at State Vocational High Schools in Batam. 12*(8), 268–277. https://doi.org/10.29303/jppipa.v12i8.15780

Labadze, L., Grigolia, M., & Machaidze, L. (2024). Role of AI chatbots in education : systematic literature review. *International Journal of Educational Technology in Higher Education*, *2023*, 1–17. https://doi.org/10.1186/s41239-023-00426-1

Li, C., Cui, H., & Serra, L. (2026). *Computers and Education : Artificial Intelligence The cognitive impact of ChatGPT in higher education : A systematic review of critical and creative thinking outcomes. 10*(January).

Mardiana, A. L., Arwizet, K., Mardizal, J., Habib, A., Desky, A., & Mardiana, A. L. (2026). *Development of a Job Sheet for Mechanical Measuring Instruments Using Virtual Learning and Project-Based Learning in the Motorcycle Engineering Department at SMK Negeri 2 Palembang. 12*(5), 105–117. https://doi.org/10.29303/jppipa.v12i5.15136

Masriani, M., Syah, N., Waskito, W., Rukun, K., Habib, A., & Desky, A. (2026). *Lentera Negeri Development of applied physics instructional materials for the technology and engineering specialization. 7*(1), 1389–1402.

Muh, A., Karmila, F., & Khalifah, M. (2025). *The efficacy of Pj4CS (integrated project- based and 4C-scaffolding) on higher-order thinking skills Abstract : 11*(1), 1–9.

Ningsih, W., Zen, Z., J, F. Y., Mirshad, E., & Arrasyidi, A. H. (2025). *The Effect of the Application of BSCS 5E Instructional Design on Communication Skills and Student Learning Outcomes in Physics Subjects at SMAN 16 Batam. 11*(10), 372–381.

Pebrianti, L., Giatman, M., Habib, A., & Desky, A. (2026). *Effects of Project-Based Learning (PjBL) Model on Learning Outcomes in Electrical Engineering Drawing. 12*(8), 278–288. https://doi.org/10.29303/jppipa.v12i8.15830

Risna, M., Arwizet, K., Habib, A., & Desky, A. (2026). *Development of an Electronic Job Sheet for Forward-Reverse Circuits in The Electrical Installation Technology Program. 12*(5), 286–295.

https://doi.org/10.29303/jppipa.v12i5.15267
Sari, N., Wulansari, R. E., Dewi, I. P., Habib, A., Desky, A., & Sari, N. (2026). *The Effect of Android-Based Learning and Learning Motivation on Learning Outcomes in Computer Science in the 10th Grade at SMK Negeri 1 Sumatera Barat*. *12*(6), 260–266. https://doi.org/10.29303/jppipa.v12i6.15001
Sudianti, N., J, F. Y., Zen, Z., Mirshad, E., Arrasyidi, A. H., & Sudianti, N. (2025). *The Effectiveness of Using Chemistry Literacy-Based Learning Videos on the Problem-Solving Ability of Grade X Students at SMA Negeri 16 Batam*. *11*(10), 382–388. https://doi.org/10.29303/jppipa.v11i10.12182
Taali, Desky, A. H. A., Pratama, A. J., & Setiawan, H. (2024). Effectiveness of Blended Learning Model in Microprocessor Course with Google Classroom. *Jurnal Penelitian Pendidikan IPA*, *10*(7), 3674–3680. https://doi.org/10.29303/jppipa.v10i7.7754
Tananda, O., Purwanto, W., Maksum, H., Giatman, M., Habib, A., & Desky, A. (2026). *Development of a Flipbook-Based E-Module with Project-Based Learning (PjBL) Model in the Subject of Ship Outfit Construction (3D Modelling) at SMKN 5 Batam*. *12*(5), 953–959. https://doi.org/10.29303/jppipa.v12i5.15035
Vaesar, S. S., Giatman, M., Rizal, F., Habib, A., & Desky, A. (2026). *The Effect of PjBL Model Assisted by Revit Application on Students ' Building Drawing Learning Outcomes in the DPIB Department at SMK Negeri 4 Palembang*. *12*(5), 183–193. https://doi.org/10.29303/jppipa.v12i5.15181
Virnindo, A., Yustisia, H., Hidayat, H., Hidayat, N., Habib, A., & Desky, A. (2026). *Evaluation of Competency-Based Learning at Vocational High School Using the CIPP Model*. *12*(5), 104–113. https://doi.org/10.29303/jppipa.v12i8.15824
Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). *ChatGPT's impact on student learning outcomes: a meta-analysis of 35 experimental studies*. 1–17.
Zhang, T., Lai, Y. C., Leung, P., & Yu, H. (2026). *Generative artificial intelligence in K-12 education : A systematic review*.
Zhao, Y., Yue, Y., Sun, Z., Jiang, Q., & Li, G. (2025). *Does Generative Artificial Intelligence Improve Students ' Higher-Order Thinking ? A Meta-Analysis Based on 29 Experiments and Quasi-Experiments*. 1–21.