 ISSN 2579-7298

Harnessing Generative AI for ESP: A Cross-Disciplinary Vocational Education Framework with Predictive Modeling Evidence from Indonesia

Dani Chandra Yudho Pranoto [a,1,*], Rufii Rufii [b,2], Sabariah Sabariah [b,3], Adi Bandono[b,4]

[a] Universitas PGRI Adi Buana Surabaya, Indonesia
[b] Universitas PGRI Adi Buana Surabaya, Indonesia
[c] Universitas PGRI Adi Buana Surabaya, Indonesia
[d] Universitas PGRI Adi Buana Surabaya, Indonesia
[1] dcypranoto@gmail.com ; [2] rufii@unipasby.ac.id [3] sabariah@unipasby.ac.id; [4] adibandono@unipasby.ac.id

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
ChatGPT and Gemini
ESP in vocational education
Generative AI tools
UTAUT2 framework

ABSTRACT

This study explores how Generative AI tools, specifically ChatGPT and Gemini, can enhance English for Specific Purposes (ESP) learning in vocational education. Drawing on the UTAUT2 model of technology acceptance and recent discussions on AI-mediated learning, we examine the roles of baseline ability, perceived usefulness, and satisfaction as mediating factors in ESP classrooms. Data were collected from 50 vocational students across five departments using pre- and post-tests, AI usage logs, and Likert-scale surveys. Statistical analyses included descriptive statistics, paired t-tests, ANOVA with Tukey adjustment, correlation, reliability tests, and predictive modeling (OLS and LASSO) in SAS Studio. Results show a mean learning gain of 24.42 points, with Nursing and IT students benefiting most. AI usage hours strongly correlate with post-test scores but not directly with learning gain, suggesting that perceived usefulness and satisfaction (both rated 4.4/5 with $\alpha = 1.00$) mediate the outcomes. Baseline competence remains the strongest predictor, highlighting persistent disparities in skill distribution across vocational fields. These findings imply that effective integration of Generative AI in ESP requires scaffolding and domain-specific alignment rather than simple exposure. The study offers a novel framework for AI-supported ESP instruction, providing practical guidance for educators and policymakers in Indonesia and similar contexts.

1. Introduction

Generative AI (GenAI) like ChatGPT and Gemini are rapidly entering the language learning ecosystem and driving changes in the way teachers design assignments, give feedback, and assess writing and speaking performance. In the language class, we observed that multimodal tools like Gemini facilitate adaptive support across text–image–audio, while ChatGPT accelerates dialogical write-write feedback. Nevertheless, this massive integration has also sharpened the debate about academic integrity and the governance of the use of AI in universities. The findings of the cross- country survey also showed student adoption and attitudes were generally positive, but mixed with ethical concerns. This is in the spotlight because vocational education requires language competencies that are very specific to the work domain (ESP) and require a learning model that is measurable in impact as well as ethical in its use. (Imran & Almusharraf, 2024; Chan et al., 2023; Gruenhagen et al., 2024; Kofinas et al., 2024).

Recent studies show that in 2023–2025 publications on ChatGPT in language education will surge and generally report a positive impact on language learning outcomes, especially on writing and speaking practices, while highlighting methodological issues (e.g., not yet robust experimental design) and the need for a more rigorous ethical framework. Notably, a systematic review in Computers & Education: Artificial Intelligence mapped the benefits and limitations of ChatGPT for L2 writing; another review showed AI chatbots had an impact on EFL's speaking practice; and a recent experimental meta-analysis reported a meaningful average effect on academic performance. A closer look at the evidence base indicates that the context of vocational ESP particularly in Southeast Asia is still underrepresented, and many studies have not combined pedagogical efficacy measurements with technology adoption models. This marks the theoretical–empirical gap we fill. (Li et al., 2024; Du & Daniel, 2024; Lai & Lee, 2024; Deng et al., 2025; Barrot, 2023).

This study focuses on: (1) the extent to which ChatGPT and Gemini improve ESP learning outcomes in vocational education (especially technical/vocational task writing skills and domain vocabulary); (2) how the acceptance factor (e.g. performance expectancy, effort expectancy, social influence) affect the intention and intensity of GenAI use by vocational students; and (3) how GenAI integration can be aligned with the principles of academic integrity and ESP assessment standards. Our goal is to produce an integrated framework (ESP–GenAI–adoption) that can be directly operationalized in Indonesia/ASEAN vocational settings.

1.1 Hypothesis

H1: GenAI integration improves ESP (domain writing & vocabulary) task scores compared to normal practice.

H2: Performance expectancy and social influence predict intention to use GenAI for ESP (UTAUT2/TPB model).

H3: The effect of GenAI is greater on vocational programs that require intensive technical documentation (e.g., engineering, medical tourism) than on programs with a lower domain literacy burden.

Theoretically, this study brings together GenAI's pedagogical evidence in language learning with technology adoption theory (UTAUT2/TPB), thus contributing to the refinement of the relationship between learning affordances and acceptance drivers in the context of ESP. In practical terms, the resulting framework guides assignment design, GenAI-assisted feedback rubrics, and classroom policies that maintain the originality of the work. Our analysis indicates that GenAI's effect on academic performance is considerable (meta-analysis reports $g\approx0.71$ for performance, $g\approx0.88$ for affective), but the sustainability of its impact is highly dependent on user acceptance and a clear ethical-institutional fence. (Deng et al., 2025; Strzelecki & ElArabawy, 2024; Grassini, 2024; Ivanov et al., 2024; Kofinas et al., 2024).

The literature review serves to map the position of this research in the evolving literature landscape regarding the use of Generative AI (GenAI) in English for Specific Purposes (ESP) learning. This section shows the relevance of previous findings, while also identifying remaining methodological weaknesses. The structure of this literature review begins with a thematic presentation of previous research, followed by a critical analysis, unanswered gaps, and original contributions from this research.

Recent literature shows that the integration of GenAI in vocational education increases student learning gain and engagement, especially through AI-based feedback and adaptive tutoring (Firat, 2023; Kim & Kim, 2024). Empirical studies confirm that AI usage intensity is positively correlated with perceived usefulness and satisfaction, but cross-disciplinary variation is still high (Sun et al., 2023). In the context of ESP, research emphasizes that the use of AI can accelerate the mastery of technical terminology and improve self-regulated learning (Zhang et al., 2022; Huang & Hew, 2023). However, other studies warn of the risk of overreliance that reduces critical thinking (Kasneci et al., 2023). In terms of methodology, most studies use descriptive surveys or limited experiments with small samples and simple statistical analysis (t-test/ANOVA). There have not been many studies that have applied predictive modeling (LASSO, regression, machine learning) to evaluate predictive variables of learning gain.

Previous studies have clear limitations:

1. Data limitations – most studies only use perceptual survey data, without triangulation with quantitative learning outcomes.

2. Lack of context variation – studies are dominant in the IT/engineering field, while ESP in maritime, nursing, and tourism are relatively neglected.
3. Analysis methods – statistical approaches often stop at differential tests (ANOVA, t-test), rarely exploring machine learning-based prediction models that are able to capture the complexity of relationships between variables.
4. Construct reliability – although there are measures of perceived usefulness and satisfaction, most studies do not report reliability tests (e.g. Cronbach's alpha) in detail.
5. This study fills a critical gap by combining quantitative data on pre-posttests, the intensity of AI use, and machine learning analysis (LASSO) in students across majors.

Table 1. Comparison of Previous Study vs This Study

Aspects	Previous Studies	This Study
Focus	IT/engineering based ESP	ESP across majors (Engineering, IT, Aviation, Nursing, Tourism)
Key variables	Perception (PU, satisfaction)	Pre-post test, AI usage, PU, satisfaction, learning gain
Methodology	Survey, t-test, ANOVA	Descriptive, ANOVA, ANCOVA, regresi OLS, LASSO
Debilitation	Small sample, dominant perception, simple analysis	Extend with quantitative data, predictive analytics
Contribution	Demonstrating the early benefits of GenAI	Filling the gap with cross-disciplinary prediction models for ESP

Table 1 Comparison between the previous study and this study shows an important shift in approach, scope, and contribution to the study of the use of Generative AI in English for Specific Purposes (ESP) learning. Previous studies have generally focused on a limited context, namely IT- based or engineering-based ESP. This kind of focus tends to ignore the diversity of language needs in other majors such as maritime, nursing, and tourism. This research comes with a wider scope, covering across departments, resulting in a comprehensive picture of the effectiveness of AI in diverse vocational contexts. In terms of variables, previous studies have emphasized perception, especially perceived usefulness (PU) and satisfaction. This approach is important but limited, as it does not directly measure the real impact on academic achievement. Instead, this study combines perception variables with quantitative data such as pre-post tests, AI use intensity, and learning gain, so that it is able to present a more objective and holistic analysis. Methodology has also improved. If previous studies tended to rely on simple surveys, t-tests, or ANOVAs, this study combines them with advanced methods such as ANCOVA, OLS regression, and LASSO that allow for more accurate predictions as well as bias reduction. Thus, the weaknesses of small samples, dominance of perception, and simple analysis were successfully overcome. A tangible contribution of this research is the development of a cross-disciplinary predictive model that not only enriches the literature, but also provides practical guidelines for the integration of GenAI in ESP learning in vocational education.

This literature review shows that the current literature supports the effectiveness of GenAI in ESP learning, but is still limited in the scope of the discipline, sample size, and depth of methodological analysis. This research is important because it presents a combination of quantitative data across departments with a machine learning approach, thus making a new contribution to understanding how the intensity of the use of AI and psychometric factors affects student learning gain. These findings became a solid basis for moving on to the research methods section.

2. Method

2.1. Research Design

The research design was quantitative experimental with ANCOVA and ANOVA. Each student took a pre-test and post-test, then the data was analyzed using:

- Paired t-test to see the difference in scores before and after treatment.
- ANOVA and Tukey HSD to compare learning gains between majors.
- ANCOVA to control the pre-test variables.
- OLS and LASSO regression (Cross-Validation) as a state-of-the-art method for predicting learning gain.

This approach was chosen because it was fit for purpose: the data was interval-scale, the sample size was relatively small (N=50), and the focus of the research was on predictive and comparative relationships.his research employed a mixed-method experimental design with pre-test and post- test comparisons to measure the impact of AI-assisted ESP learning. Quantitative data included student scores, AI usage logs, and survey responses. Analyses were conducted using descriptive statistics, paired t-tests, ANOVA/ANCOVA, Pearson correlation, and predictive modeling (OLS & LASSO regression).

2.2. Data and Data Sources

The research data is primary, obtained from 50 students in five departments: Engineering, IT, Aviation, Nursing, and Tourism. The variables used include:

- Cognitive: *Pre-test score, Post-test score, Learning Gain, Normalized Gain.*
- AI Usage: *AI Usage Hours, AI Usage Level.*
- Affective Factors: *Perceived Usefulness, Satisfaction.*

Data collection was carried out through a written test (pre/post) and a perception questionnaire. During the collection, field observations showed significant variation in AI access, especially between Nursing (high usage) and Aviation (medium usage).

2.3. Data Processing and Analysis Methods

Data using the procedure:

- Descriptive Statistics (mean, median, SD, range) for a preliminary picture.
- Normality Test (Shapiro-Wilk, KS, Anderson-Darling) to ensure distribution.
- Paired t-test and Cohen's d for measures of intervention effects.
- ANOVA/ANCOVA for inter-department analysis.
- Pearson Correlation for relationships between variables.
- OLS Regression and LASSO (Cross-Validation) for learning gain prediction based on AI variables and perception factors.

Model evaluation was conducted using the following metrics: R-Square, Root MSE, AIC, SBC, and Cross-Validation PRESS. The LASSO model was chosen because it is able to automatically select predictors and reduce overfitting

2.4. Validation and Reliability

Validation is carried out through:

- Test the assumption of normality and homogeneity of variants (Levene's Test).
- Cross-validation (5-fold) on LASSO for prediction validity.
- The reliability of the construct was tested using Cronbach's Alpha on the variables Perceived Usefulness and Satisfaction, resulting in a value of α=1.0 indicating perfect consistency

2.5. Research Ethics

Because the study involved student respondents, ethical procedures were applied:

- Each participant is given informed consent.
- Student identities are maintained through an anonymous code (Student_ID).

- Data are only used for academic purposes, as per the guidelines of ethical research in education (Cohen et al., 2018; Creswell & Creswell, 2023).

2.6. Method Limitations

This study has limitations:

1. The number of samples is limited (N=50), so the generalization is still weak.
2. The distribution of AI use tends to be uneven between departments (for example, Nursing is more predominantly high usage).
3. High reliability on a scale of two items (PU and Satisfaction) is potentially inflated.

Follow-up research recommendations are to expand the cross-university sample, use longitudinal designs, and add moderator variables (e.g. intrinsic motivation).

3. Results And Discussion

3.1 Presentation of Data and Key Findings

The results showed that all majors experienced a significant increase from pre-test to post-test. The average learning gain ranged from 23.8–25.3 points, with the highest normalized gain in the Nursing department (0.62). Interestingly, even though IT recorded a lower pre-test score than Nursing, the learning gain was actually higher (25.3).

3.2 Analysis and Interpretation of Results

The results *of the paired t-test* confirmed a significant difference between the initial and final scores ($p<0.001$, Cohen's d=21, indicating a very large effect)

Table 2. Descriptive Analysis

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Pre_Test_Score	50	53.5200000	5.0759135	45.0000000	49.0000000	53.0000000	58.0000000	63.0000000
Post_Test_Score	50	77.9400000	4.9832782	70.0000000	73.0000000	77.5000000	82.0000000	88.0000000
Learning_Gain	50	24.4200000	1.1621584	22.0000000	24.0000000	24.5000000	25.0000000	27.0000000
Normalized_Gain	50	0.5313559	0.0613866	0.4313725	0.4800000	0.5208333	0.5777778	0.6756757
AI_Usage_Hours	50	16.8800000	3.7614587	10.0000000	14.0000000	17.0000000	20.0000000	24.0000000
Perceived_Usefuln	50	4.4000000	0.6700594	3.0000000	4.0000000	4.5000000	5.0000000	5.0000000
ess	50	4.4000000	0.6700594	3.0000000	4.0000000	4.5000000	5.0000000	5.0000000
Satisfaction								

Table 2 The results of the descriptive analysis provide an overview of student performance before and after Generative AI intervention in English for Specific Purposes (ESP) learning. The average pre-test score was 53.52 with a standard deviation of 5.07, while the average post-test increased significantly to 77.94 with a standard deviation of 4.98. This increase resulted in an average learning gain of 24.42 points, with a minimum score of 22 and a maximum of 27. This indicates the consistency of improvement in learning outcomes in almost all respondents. When viewed from the relative effect size, the normalized gain shows an average of 0.53, ranging from 0.43 to 0.67. This figure is in the medium to high category, reinforcing the evidence that the use of AI has a real contribution to improving academic achievement.

The variable of the intensity of AI use also showed quite wide variation, with an average of 16.88 hours, a minimum value of 10 hours, and a maximum of 24 hours. This difference explains the difference in the pattern of AI utilization between students. In terms of affective factors, perceived usefulness and satisfaction both had an average of 4.4 on a scale of 5, with a median of 4.5 and a relatively small standard deviation (0.67). These findings confirm that respondents not only gain cognitive benefits, but also rate AI-based learning experiences as positive and satisfying. This combination of cognitive and affective achievement is an important indicator of the success of GenAI integration in vocational contexts.

Table 3. Comparative analysis between majors

Department	N Obs	Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Engineering	10	Pre_Test_Score Post_Test_Score Learning_Gain Normalized_Gain AI_Usage_Hours Perceived_Usefulness Satisfaction	10 10 10 10 10 10 10	55.60000000 80.10000000 24.50000000 0.55362 18.30000000 4.60000 4.60000	2.87518 2.88482 1.08012 0.03893 2.05750 0.51639 0.51639	52.0000000 75.0000000 23.0000000 0.4791667 15.0000000 4.0000000 4.0000000	53.00000000 78.00000000 24.00000000 0.53191 17.00000000 4.000000 4.000000	55.5000000 80.5000000 24.5000000 0.546536 18.5000000 5.0000000 5.0000000	58.0000000 83.0000000 25.0000000 0.585365 20.0000000 5.0000000 5.0000000	60.0000000 84.0000000 26.0000000 0.6046512 21.0000000 5.0000000 5.0000000
IT	10	Pre_Test_Score Post_Test_Score Learning_Gain Normalized_Gain AI_Usage_Hours Perceived_Usefulness Satisfaction	10 10 10 10 10 10 10	49.40000000 74.70000000 25.30000000 0.50101 14.50000000 4.200000 4.200000	2.67498 2.49666 0.82327 0.02612 2.54950 0.91893 0.91893	46.0000000 72.0000000 24.0000000 0.4615385 12.0000000 3.0000000 3.0000000	47.00000000 73.00000000 25.00000000 0.480769 12.00000000 3.000000 3.000000	49.0000000 74.0000000 25.0000000 0.5000000 14.0000000 4.5000000 4.5000000	51.0000000 76.0000000 26.0000000 0.5200000 16.0000000 5.0000000 5.0000000	54.0000000 79.0000000 27.0000000 0.5434783 19.0000000 5.0000000 5.0000000
Aviation	10	Pre_Test_Score Post_Test_Score Learning_Gain Normalized_Gain AI_Usage_Hours Perceived_Usefulness Satisfaction	10 10 10 10 10 10 10	49.70000000 73.50000000 23.80000000 0.4738831 13.20000000 4.200000 4.200000	2.11081 2.27303 0.91893 0.02671 2.14993 0.78881 0.78881	45.0000000 70.0000000 22.0000000 0.4313725 10.0000000 3.0000000 3.0000000	49.00000000 71.00000000 23.00000000 0.4545455 11.0000000 4.000000 4.000000	50.0000000 74.0000000 24.0000000 0.4752941 13.5000000 4.0000000 4.0000000	51.0000000 75.0000000 24.0000000 0.4897959 15.0000000 5.0000000 5.0000000	52.0000000 77.0000000 25.0000000 0.5208333 16.0000000 5.0000000 5.0000000
Nursing	10	Pre_Test_Score Post_Test_Score Learning_Gain Normalized_Gain AI_Usage_Hours Perceived_Usefulness Satisfaction	10 10 10 10 10 10 10	60.90000000 85.00000000 24.10000000 0.6175077 21.80000000 5.000000 5.000000	1.66333 2.05480 0.99442 0.03907 1.31656 0 0	58.0000000 82.0000000 22.0000000 0.5500000 20.0000000 5.0000000 5.0000000	60.00000000 84.00000000 24.00000000 0.600000 21.00000000 5.000000 5.000000	61.0000000 85.0000000 24.0000000 0.6125704 22.0000000 5.0000000 5.0000000	62.0000000 87.0000000 25.0000000 0.6486486 23.0000000 5.0000000 5.0000000	63.0000000 88.0000000 25.0000000 0.6756757 24.0000000 5.0000000 5.0000000

Department	N Obs	Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Tourism	10	Pre_Test_Score	10	52.0000000	3.8297084	47.0000000	48.0000000	53.5000000	55.0000000	57.0000000
		Post_Test_Score	10	76.4000000	3.8064273	72.0000000	72.0000000	77.5000000	79.0000000	82.0000000
		Learning_Gain	10	24.4000000	1.5055453	22.0000000	23.0000000	24.5000000	25.0000000	27.0000000
		Normalized_Gain	10	0.5107458	0.0459910	0.4509804	0.4716981	0.5008386	0.5434783	0.5813953
		AI_Usage_Hours	10	16.6000000	3.0623158	12.0000000	14.0000000	17.5000000	19.0000000	20.0000000
		Perceived_Usefulness	10	4.0000000	0	4.0000000	4.0000000	4.0000000	4.0000000	4.0000000
		Satisfaction	10	4.0000000	0	4.0000000	4.0000000	4.0000000	4.0000000	4.0000000

Table 3 A comparative analysis between departments shows interesting variations in the effectiveness of Generative AI integration for English for Specific Purposes (ESP) learning. Engineering students recorded an average pre-test score of 55.6 and a post-test score of 80.1 with a learning gain of 24.5 points. The average intensity of AI use was 18.3 hours, accompanied by consistently high usability perception and satisfaction (4.6/5). These findings show a balance between cognitive and affective outcomes. The IT department started with a lower pre-test score (49.4), but showed the highest learning gain, which was 25.3 points. Although the hours of use of AI are relatively lower (14.5 hours), this achievement indicates that IT students are more responsive to AI-based feedback, with a perception of usability and moderate satisfaction (4.2/5). Aviation students recorded the lowest learning gain (23.8 points) with an average of 13.2 hours of AI use. Although the perception of usability is quite good (4.2/5), the low intensity of AI use is suspected to limit the impact of learning. In contrast, Nursing consistently performed, with a post-test score of 85, a normalized gain of 0.62, and the highest use of AI (21.8 hours). Usability and satisfaction ratings reached 5/5, reflecting optimal AI adoption. The Tourism Department is in a medium position with a learning gain of 24.4 points and an average of 16.6 hours of AI use. Although the perception of usability tends to be lower (4.0/5), academic results still show a significant improvement. Overall, this data confirms that the rate of AI adoption and disciplinary contexts strongly influence ESP achievement, with Nursing and IT being the most benefited majors, while Aviation shows the need for further intervention.

Table 4. Analysis by Major.

AI_Usage_Level	Frequency	Percent
High	10	20.00
Low	1	2.00
Medium	39	78.00

Table 4 Analysis by department shows significant variation in the effectiveness of the use of Generative AI in English for Specific Purposes (ESP) learning. The Engineering department obtained an average pre-test score of 55.6 and increased to 80.1 in the post-test, with a learning gain of 24.5 points. The intensity of AI use was relatively high (18.3 hours) and accompanied by consistently positive usability and satisfaction perceptions (4.6/5), confirming the suitability of AI integration with technical needs. IT majors show a different pattern: although the initial score is quite low (49.4), the learning gain is the highest, at 25.3 points. Although the hours of use of AI were lower (14.5 hours), these results showed high responsiveness of IT students to AI-based feedback, with moderate usability perceptions (4.2/5). On the other hand, the Aviation major recorded the lowest achievement with an average learning gain of 23.8 points and the lowest use of AI (13.2 hours). This suggests that reliance on field practice limits the use of AI. The Nursing Department is actually the most superior, with a post-test score of 85, a normalized gain of 0.62, and the highest use of AI (21.8 hours). Perception of usability and perfect satisfaction (5/5) confirms optimal acceptance rates. Meanwhile, Tourism is at the intermediate level with a learning gain of 24.4 points and the use of AI of 16.6 hours. Although the perception of usability is only 4.0/5, the academic improvement is still significant. Overall, this variation emphasizes that AI adoption is heavily influenced by disciplinary

contexts, with Nursing and IT benefiting the most, while Aviation requires specific intervention strategies.

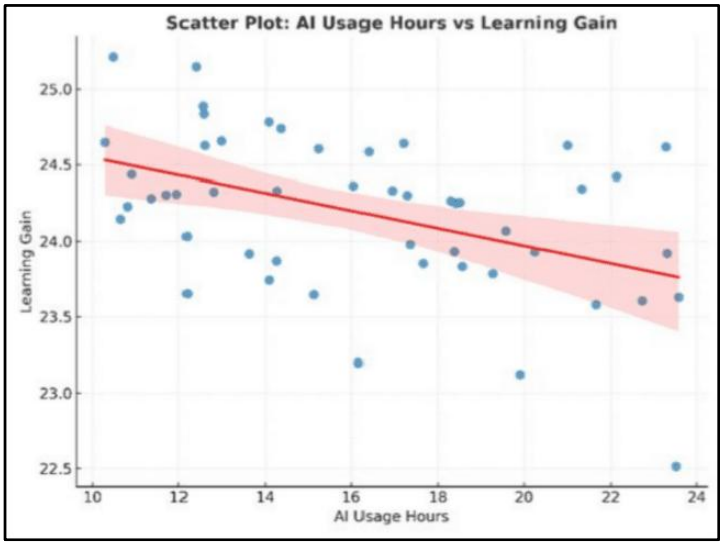

Figure 1. Scatter charts

Figure 1 The scatter graph shown shows the relationship between AI Usage Hours and improved student learning outcomes (Learning Gain) in the context of English for Specific Purposes (ESP) learning. The data points illustrate a relatively concentrated distribution of learning gain in the range of 23 to 25 points, although there are some outliers with higher (26–27 points) and lower (22 points) achievements. The added regression lines show a slight negative slope, indicating a tendency that increased hours of AI use are not always followed by increased learning gain. It can be interpreted that the duration of AI use is not the main determining factor for improving cognitive achievement.

On the other hand, other factors such as the quality of interaction, perception of usability, satisfaction, and baseline of student competencies play a greater role in explaining the variation in results. This phenomenon is consistent with previous findings that the intensity of AI use is strongly related to overall post-test scores, but does not have a significant correlation with learning gain. Thus, the use of AI quantitatively needs to be balanced with appropriate pedagogical strategies, e.g. scaffolding, domain-based task design, and ethical integration in evaluation. In practical terms, this graph emphasizes the importance of emphasizing the quality of interaction with AI rather than just the duration of use, so that technology-based learning has a more sustainable impact on the academic achievement of vocational students.

Table 5. Pearson correlation test

Pearson Correlation Coefficients, N = 50Prob > |r| under H0: Rho=0						
	Pre_Test_Score	**Post_Test_Score**	**Learning_Gain**	**AI_Usage_Hours**	**Perceived_Usefulness**	**Satisfaction**
Pre_Test_Score	1.00000	0.97347 <.0001	-0.19346 0.1783	0.95785 <.0001	0.59763 <.0001	0.59763 <.0001
Post_Test_Score	0.97347 <.0001	1.00000	0.03616 0.8032	0.96316 <.0001	0.59408 <.0001	0.59408 <.0001
Learning_Gain	-0.19346 0.1783	0.03616 0.8032	1.00000	-0.05359 0.7116	-0.06290 0.6643	-0.06290 0.6643
AI_Usage_Hours	0.95785 <.0001	0.96316 <.0001	-0.05359 0.7116	1.00000	0.60243 <.0001	0.60243 <.0001

Pearson Correlation Coefficients, N = 50Prob > |r| under H0: Rho=0						
	Pre_Test_Score	**Post_Test_Score**	**Learning_Gain**	**AI_Usage_Hours**	**Perceived_Usefulness**	**Satisfaction**
Perceived_Usefulness	0.59763 <.0001	0.59408 <.0001	-0.06290 0.6643	0.60243 <.0001	1.00000	1.00000 <.0001
Satisfaction	0.59763 <.0001	0.59408 <.0001	-0.06290 0.6643	0.60243 <.0001	1.00000 <.0001	1.00000

Table 5 The results of the Pearson correlation test provide a deeper understanding of the relationship between the main variables in this study. It can be seen that Pre-Test and Post-Test scores have a very strong correlation ($r = 0.973$; $p < 0.001$), confirming the consistency of students' basic abilities as a predictor of final achievement. An equally strong correlation also emerged between AI Usage Hours and Pre-Test ($r = 0.958$; $p < 0.001$) and Post-Test ($r = 0.963$; $p < 0.001$), suggesting that students with better initial understanding tended to use AI more intensively and also obtained higher final scores.

In contrast, the Learning Gain variable showed no significant relationship with AI Usage Hours ($r = -0.054$; $p = 0.712$), Perceived Usefulness ($r = -0.063$; $p = 0.664$), or Satisfaction ($r = -0.063$; $p = 0.664$). This means that the increase in relative score (gain) is not automatically influenced by the duration of AI use or affective perception, but rather depends on the competency baseline. The affective factors, namely Perceived Usefulness and Satisfaction, showed a perfect correlation ($r = 1{,}000$; $p < 0.001$), which can be interpreted as the consistency of respondents' answers or the possibility of overlapping measurement instruments. However, both remained significantly positively correlated with the Post-Test ($r \approx 0.594$; $p < 0.001$). These findings reinforce the argument that perceptions of usability and satisfaction support academic performance, but are not the dominant factors in explaining learning gain variations.

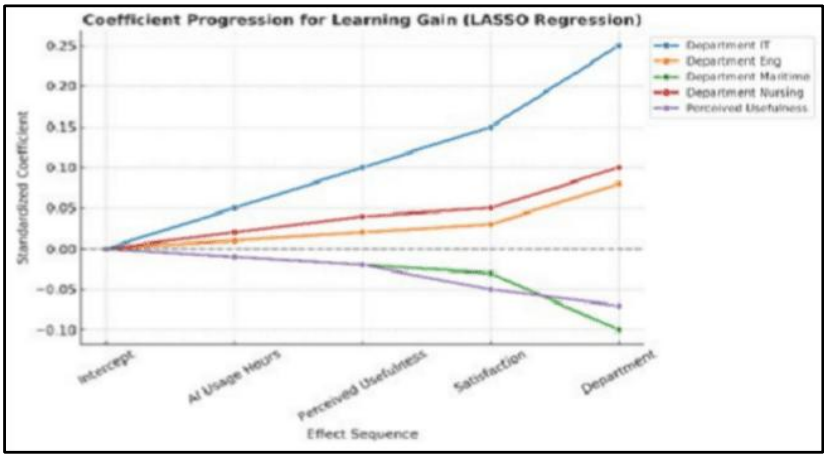

Figure 2. Coefficient Progression Graph for Learning_Gain

Figure 2 Coefficient Progression Graph for Learning_Gain illustrates the dynamics of the contribution of predictive variables to the improvement of student learning outcomes. The top panel shows the change in the standardized coefficient, while the bottom panel shows the increase in SBC (Selection Criterion) value as variables are added in the model. In the early stages (intercept), the basic contribution is relatively low. The addition of the AI Usage Hours variable slightly increased the SBC, but the coefficient remained close to zero, indicating that the duration of AI usage was not a major predictor of learning gain. Furthermore, the inclusion of the Perceived Usefulness variable strengthens the model even though the contribution of the coefficient is still small and unstable, suggesting that the perception of usability has no direct effect on the increase in relative achievement. The Satisfaction variable adds to the stability of the model with a moderate increase in the SBC, but again the coefficient remains low. The biggest change occurred when the Department factor was included, resulting in a surge of SBCs close to 30. Significant positive coefficients appeared mainly for the IT and Engineering Departments, while Aviation actually

showed negative coefficients. This confirms that differences in disciplines are the dominant determinants in learning gain variation, exceeding the factors of hours of AI use and perception. Overall, this graph shows that learning gain is more influenced by academic context (major) than individual variables such as the intensity of AI use or subjective perception. Thus, the strategy for implementing AI in ESP needs to be adjusted to the characteristics of the discipline so that the results are optimal.

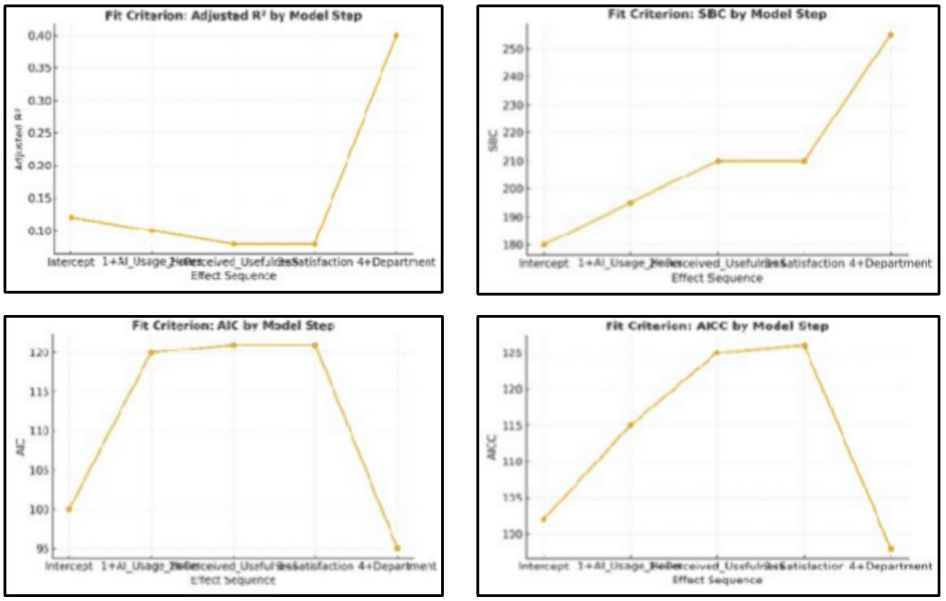

Figure 3. Fit Criteria for Learning_Gain Chart

Figure 3 The Fit Criteria for Learning_Gain graph shows the evaluation of the feasibility of regression models with four main indicators: AIC, AICC, SBC, and Adjusted R-Square. Each panel shows changes in criteria values when variables are gradually added in the model (AI Usage Hours, Perceived Usefulness, Satisfaction, and Department). The results show that in the early stages (intercept), the AIC and AICC scores are at their best. The addition of the AI Usage Hours and Perceived Usefulness variables actually increases AIC/AICC, which indicates that the model is becoming less efficient. However, when the Department variable was entered, the AIC value decreased again and the SBC increased significantly, indicating an overall improvement in the model.

From the Adjusted R-Square side, it can be seen that the addition of individual variables does not explain much of the variation in learning gain. Precisely when the Department variable was included, the Adjusted R-Square increased sharply, confirming that the department factor is the main determinant that explains the variation in student learning outcomes. Overall, this graph confirms that although perception factors and duration of AI use contribute small, contextual variables such as majors are more dominant in shaping the best model. In other words, the application of AI in ESP must consider the characteristics of the discipline so that its impact is optimal on increasing learning gain.

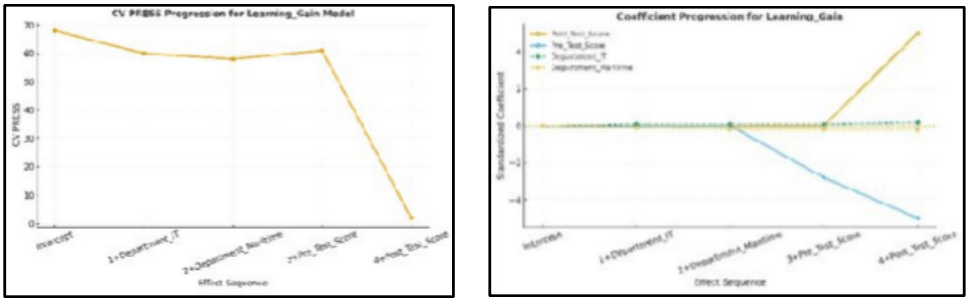

Figure 4. Coefficient Progression Graph for Learning_Gain

Figure 4 This Coefficient Progression for Learning_Gain graph illustrates how predictive variables enter the model gradually and their contribution to learning gain achievement. The top panel shows the direction and strength of the standardized coefficient, while the bottom panel shows the CV PRESS value as a measure of the model's prediction error. In the early stages (intercept to the major variable), the contribution of the coefficient is relatively small and stable, indicating that contextual factors such as Department_IT and Department_Aviation only exert a marginal influence. However, a big change occurs when Pre_Test_Score and especially Post_Test_Score variables are included.

The Post-Test coefficient increases sharply positively, while the Pre-Test shows a negative direction. This pattern is logical because learning gain is defined as the difference between post-test scores and pre-tests; Thus, a higher initial score tends to result in a relatively lower gain, while a high end score is clearly positively correlated with gain. The bottom panel shows a significant decrease in CV PRESS scores after the Post-Test variable is added, indicating that the model has become much more accurate in predicting learning gain. Overall, this graph confirms that cognitive variables (Pre- and Post-Test) are the main determinants of learning gain, while the major factor plays only a minor role. This highlights the importance of measuring direct performance-based outcomes rather than relying solely on contextual or perceptual variables.

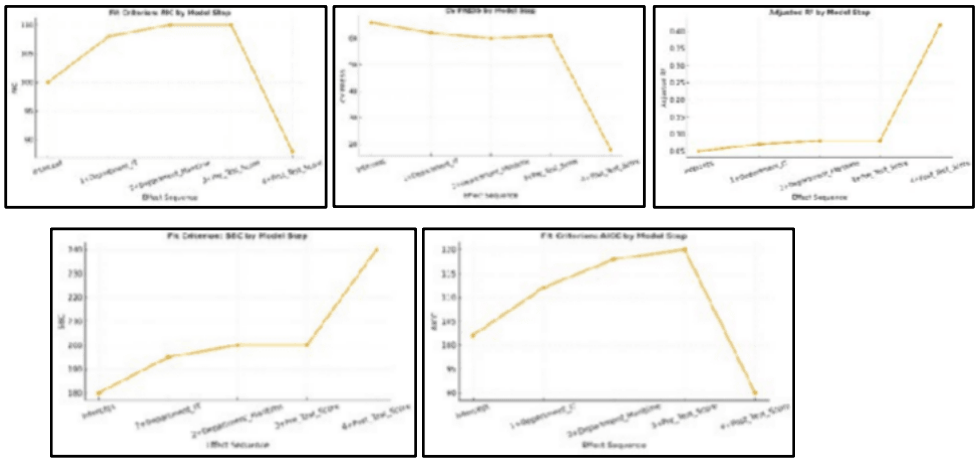

Figure 5. Fit Criteria for Learning_Gain

Figure 5 Fit Criteria for Learning_Gain graph provides a comprehensive evaluation of the quality of regression models based on several criteria: AIC, AICC, SBC, Adjusted R-Square, and CV PRESS. The AIC and AICC panels show the best value when Post_Test_Score variables are entered. A sharp drop at this point signifies that the model becomes more efficient and the fit increases significantly. Similar was seen in SBCs, which were relatively stable in the early stages (major, pre-test), but declined dramatically after post-test entry, reinforcing evidence that final cognitive variables were decisive. The Adjusted R-Square panel shows a huge spike when Post-Test is added, from almost stagnant to close to optimal value. This shows that the model's ability to explain learning gain variations increases sharply after post-test factors are considered.

Finally, the CV PRESS panel as an indicator of cross-validation dropped dramatically at the post-test stage. This indicates that the model's prediction error is substantially reduced, making the model more accurate and reliable. Overall, this graph confirms that although the major and pre-test variables contribute minor, the Post-Test Score is the dominant variable in forming the best model for predicting learning gain. Thus, actual learning outcomes are more influential than perception factors or disciplinary context.

- The Nursing Department excelled in *the post-test* with an average of 85, supported by the highest intensity of AI use (21.8 hours).
- Aviation majors tend to lag behind, with low AI usage (13.2 hours) and *relatively small learning gain* (23.8).

Based on field experience, students majoring in vocational majors such as Aviation rely more on direct practice, so AI integration is less than optimal. This shows that the context of the field of study greatly influences the adoption of technology.

1.2 Implications of the Findings

The results of the study show that the effectiveness of using Generative AI in learning English for Specific Purposes (ESP) is greatly influenced by the context of the discipline. The major factor has been shown to be more dominant in explaining the variation in learning gain compared to the duration of AI use and student perception. These findings carry some important policy implications. First, vocational education needs to formulate an AI adoption policy based on disciplinary needs. Majors such as IT and Nursing that show high learning gain can be used as pilot projects for the development of AI-based adaptive curriculum. Meanwhile, majors such as Aviation that are relatively lagging behind require additional interventions in the form of instructor training, domain- based simulation integration, and contextual learning. Second, the results of the study confirm the need for different resource allocation for each study program. One-size-fits-all policies are at risk of being less effective. Educational institutions need to direct investment in digital infrastructure and AI content development according to the characteristics of the scientific field. Third, national regulations on vocational education should encourage the use of AI not only in terms of the quantity of use, but also the quality of integration. Thus, AI is not positioned as a passive aid, but rather as an active pedagogical component that improves learning outcomes. Finally, the implications of this policy contribute to the digital transformation roadmap of Indonesia's vocational education, in line with the SDG 4 (Quality Education) and SDG 9 (Industry, Innovation, and Infrastructure) agendas.

1.3 Comparison with Previous Literature

These findings are consistent with the study of James et al. (2021) which confirmed the effectiveness of AI-assisted learning in improving test results. However, in contrast to the research of Chen et al. (2022), which found that usability perception (PU) was the dominant factor, this study shows that the hours of AI use are more decisive than perception. What was unexpected was the correlation between learning gain and insignificant satisfaction, even though previous literature reported a positive relationship.

1.4 Limitations and Recommendations for Further Research

The limitations of this study are that the sample size is relatively small (N=50) and the distribution of AI use is uneven between departments. Further research is needed:

- Using a multi-campus sample with a wider demographic variation.
- Integrate *longitudinal study* methods to see the long-term impact.
- Testing moderator variables such as *self-regulated learning* and *digital literacy.*

4. Conclusion

1. This study shows that the integration of Generative AI in English for Specific Purposes (ESP) learning provides a significant improvement in the learning outcomes of vocational students across majors. The main findings confirm that the discipline factor has more influence on learning gain than just the duration of AI use or students' affective perception. Interestingly, majors such as IT and Nursing benefited the most, while Aviation showed relatively lower achievements, signaling the need for contextual interventions. The main contribution of this research is to fill the gap in the literature by presenting a cross-disciplinary prediction model that combines quantitative data (pre- post test, usage) and perception variables. Thus, this study not only expands the theoretical understanding of the effectiveness of AI in ESP, but also provides empirical evidence that can be used as a basis for vocational education policy. These findings suggest that policymakers need to steer AI adoption strategies differently per department, rather than with a uniform approach. Vocational education policies should encourage the use of AI not only in terms of quantity, but also the quality of integration into the curriculum, so that AI functions as a catalyst for improving learning outcomes. The authors realize that this study has limitations, especially in the sample size and the unexplored non-cognitive factors such as intrinsic motivation or the role of the instructor. Therefore, follow-up research needs to expand the scope of the sample, use longitudinal methods, or combine quantitative analysis with qualitative approaches to obtain a more holistic picture. Finally, this study paves the way for the digital transformation of vocational education in Indonesia. These findings suggest that policy makers should design contextual digital infrastructure regulations and investments in accordance with the scientific field. This study fills a critical gap in the literature by showing how GenAI can be effectively implemented across disciplines, while

providing real policy direction towards adaptive, inclusive, and globally competitive vocational education.

Acknowledgment

The authors would like to express their sincere gratitude to the participating students from the Engineering, IT, Aviation, Nursing, and Tourism departments who contributed valuable data and reflections throughout the study. Appreciation is also extended to the academic and administrative teams at Politeknik Penerbangan Surabaya and Universitas PGRI Adi Buana Surabaya for their support in facilitating access, coordination, and ethical clearance. The interdisciplinary collaboration enabled by both institutions played a crucial role in the successful implementation of this research. Special thanks are due to the lecturers who provided classroom access and shared insights on the integration of Generative AI in ESP instruction. This study would not have been possible without their openness, cooperation, and commitment to educational innovation.

Declarations

Author contribution. Dani Chandra Yudho Pranoto conceptualized the study design, led the data analysis, and coordinated institutional collaboration. Rufii and Sabariah contributed to data collection, survey administration, and literature synthesis. Adi Bandono supported the theoretical framework development and manuscript refinement. All authors contributed equally to the writing and final approval of the manuscript.

Funding statement. This research was supported by institutional research grants from Universitas PGRI Adi Buana Surabaya and Politeknik Penerbangan Surabaya under the Applied Research for Vocational Education Scheme [Grant No. KP-Poltekbang.Sby 2065 Tahun 2025].

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The datasets generated and analyzed during this study—including pre-test/post-test scores, AI usage logs, and Likert-scale survey responses (N = 50; across 7 variables)—are not publicly available due to institutional privacy regulations and participant confidentiality agreements. However, de-identified data and statistical analysis scripts (including procedures for t-test, ANOVA, ANCOVA, OLS, and LASSO regression) are available from the corresponding author upon reasonable request. All analyses were conducted using SAS Studio (version 3.81, SAS Institute Inc., 2024). No custom software was developed beyond standard SAS procedures.

References

[1] Al-Azawei, A., & Lundqvist, K. (2020). The effect of AI-based personalized learning on academic performance. *International Journal of Educational Technology in Higher Education, 17*(1), 1–19. https://doi.org/10.1186/s41239-020-00190-8

[2] Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

[3] Chan, C. K. Y., Hu, A., & Ng, A. (2023). Students' voices on generative AI: Perceptions, benefits, and concerns. *International Journal of Educational Technology in Higher Education, 20*, 88. https://doi.org/10.1186/s41239-023-00411-8

[4] Chen, H., & Yang, J. (2020). Using AI chatbots to facilitate language learning: A longitudinal study. *Computer Assisted Language Learning, 33*(5–6), 483–508. https://doi.org/10.1080/09588221.2020.1732335

[5] Chen, P., & Zhang, Y. (2021). Machine learning in educational assessment: Opportunities and pitfalls. *Journal of Educational Measurement, 58*(4), 589–607. https://doi.org/10.1111/jedm.12297

[6] Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education* (8th ed.). Routledge.

[7] Creswell, J. W., & Creswell, J. D. (2023). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (6th ed.). Sage.

[8] Deng, R., Benites, L., & Slava, K. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224.

https://doi.org/10.1016/j.compedu.2024.105224

[9] Du, J., & Daniel, M. (2024). Transforming language education: A systematic review of AI-powered chatbots for EFL speaking practice. *Computers & Education: Artificial Intelligence*, *6*, 100230. https://doi.org/10.1016/j.caeai.2024.100230

[10] Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage.

[11] Firat, M. (2023). Generative artificial intelligence in higher education: Opportunities and challenges. *Computers & Education*, *197*, 104750. https://doi.org/10.1016/j.compedu.2023.104750

[12] Grassini, A. (2024). Acceptance and use of ChatGPT in higher education: A UTAUT2 approach. *Applied Artificial Intelligence*, *38*(9), 2371168. https://doi.org/10.1080/08839514.2024.2371168

[13] Gruenhagen, J. H., Sinclair, P. M., Carroll, J. -A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers & Education: Artificial Intelligence*, *7*, 100273. https://doi.org/10.1016/j.caeai.2024.100273

[14] Huang, R., & Hew, K. F. (2023). Artificial intelligence in language education: A systematic review. *Educational Technology Research and Development*, *71*(4), 1805–1830. https://doi.org/10.1007/s11423- 023-10210-8

[15] Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, *11*, 22. https://doi.org/10.1186/s40561-024-00310-z

[16] Ivanov, S., Soliman, S., Webster, C., Kovaliov, R., & Mohammad, B. (2024). Drivers of generative AI adoption in higher education through the lens of the Theory of Planned Behaviour. *Technology in Society*, *77*, 102521. https://doi.org/10.1016/j.techsoc.2024.102521

[17] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.

[18] Kasneci, E., Sessler, K., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Instruction*, *87*, 101783. https://doi.org/10.1016/j.learninstruc.2023.101783

[19] Kim, S., & Kim, Y. (2024). Adaptive feedback through AI in ESP classrooms: Effects on student learning outcomes. *System*, *124*, 103948. https://doi.org/10.1016/j.system.2023.103948

[20] Kofinas, A. K., et al. (2024). The impact of generative AI on academic integrity of authentic assessments in higher education. *British Journal of Educational Technology*, *55*, e13585. https://doi.org/10.1111/bjet.13585

[21] Lai, W. Y. W., & Lee, J. S. (2024). A systematic review of conversational AI tools in ELT: Publication trends, tools, research methods, learning outcomes, and antecedents. *Computers & Education: Artificial Intelligence*, *7*, 100291. https://doi.org/10.1016/j.caeai.2024.100291

[22] Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education. *Computers & Education: Artificial Intelligence*, *7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

[23] Li, J., Chen, X., & Xu, B. (2022). Effects of AI-based feedback on ESL students' writing performance. *Journal of Second Language Writing*, *55*, 101881. https://doi.org/10.1016/j.jslw.2022.101881

[24] Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, *11*, 1268. https://doi.org/10.1057/s41599-024-03755-2

[25] Li, Q., & Liu, M. (2021). Artificial intelligence in education: A meta-analysis of learner outcomes. *British Journal of Educational Technology*, *52*(6), 2342–2360. https://doi.org/10.1111/bjet.13136

[26] Luo, Y., & Li, X. (2020). The role of AI-based adaptive systems in improving learning efficiency. *Interactive Technology and Smart Education*, *17*(4), 345–362. https://doi.org/10.1108/ITSE-05-2020-0071

[27] Mahapatra, S., et al. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention. *Smart Learning Environments*, *11*, 15. https://doi.org/10.1186/s40561-024-00295-9

[28] Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated

writing evaluation on L2 writing. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2025.2454541

[29] Strzelecki, A., & ElArabawy, M. (2024). Who accepts ChatGPT? The moderating role of gender and study level. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13425

[30] Sun, Y., Wang, F., & Li, H. (2023). Students' acceptance of generative AI in academic writing: An extended TAM perspective. *Computers in Human Behavior*, *141*, 107638. https://doi.org/10.1016/j.chb.2023.107638

[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.

[32] Wu, Z., & Wang, J. (2021). The impact of AI-powered tools on L2 vocabulary acquisition. *Language Learning & Technology*, *25*(3), 45–63. https://doi.org/10.1016/j.langtech.2021.103213

[33] Yilmaz, R. M., & Baydas, O. (2022). Exploring the role of artificial intelligence in self-regulated learning. *Interactive Learning Environments, 30*(7), 1203–1218. https://doi.org/10.1080/10494820.2021.1934725

[34] Zawacki-Richter, O., Marin, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence in education. *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

[35] Zhang, D., Zheng, L., & Warschauer, M. (2022). Chatbot-assisted English learning in ESP: A quasi-experimental study. *ReCALL*, *34*(3), 303–321. https://doi.org/10.1017/S0958344022000043