

Causal Machine Learning to Estimate the Impact of Using Generative AI on Students' GPA And Skill Retention

Angelina Hadriani¹, Ade Sumaedi², Layli Ana³, Sugiyanti⁴, Yuni Yulistiani⁵

^{1,2,3,4,5}Department of Computer Systems (Serang City Campus), Universitas Pamulang, Indonesia

ABSTRACT

The rapid adoption of Generative Artificial Intelligence (GenAI) in higher education raises questions about its effects on academic performance and durable skills. This study evaluates the adjusted relationship between GenAI usage, GPA Change, and Skill Retention Score using 50,000 student records. Weekly GenAI Hours is treated as the exposure, while academic, behavioral, institutional, and psychological characteristics are included as adjustment covariates. A cross-fitted Double Machine Learning workflow is supported by overlap diagnostics, continuous-treatment sensitivity analysis, quartile comparison, and exploratory heterogeneous-effect analysis. High GenAI usage is associated with higher GPA Change (ATE = 0.0328; 95% CI: 0.0297–0.0358) and Skill Retention Score (ATE = 0.8098; 95% CI: 0.5733–1.0464). The quartile comparison shows the same direction. However, each additional weekly GenAI hour is associated with lower skill retention (ATE = –0.1096; 95% CI: –0.1298 to –0.0895). Thus, greater usage intensity does not necessarily indicate deeper learning. Because the data are observational, the estimates remain conditional on measured covariates. Universities should promote GenAI as a structured learning aid while preserving independent reasoning and sustained skill development.

Keyword : Generative Artificial Intelligence, Double Machine Learning, Causal Inference, Academic Performance, Skill Retention



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Ade Sumaedi,
Department of Computer Systems (Serang City Campus),
Universitas Pamulang,
Jalan Pantura No.6, Kalodran, Kec. Walantaka, Kota Serang, Banten 42183.
Email: adesumaedi10093@unpam.ac.id

1. INTRODUCTION

Generative Artificial Intelligence (GenAI) has moved rapidly from an experimental technology to an everyday learning resource in higher education. Large language model-based tools can provide explanations, generate examples, summarize materials, support academic writing, and assist with programming tasks. In computer-related disciplines, the same tools can also produce code, debugging suggestions, and step-by-step problem-solving guidance. These capabilities may broaden access to learning support, but they also alter the amount and type of cognitive effort students invest in academic work [1].

The educational value of GenAI is therefore not determined by access alone. UNESCO emphasizes human-centred implementation, transparency, privacy, and AI literacy, while the OECD notes that learning benefits are more likely when GenAI is embedded in clear pedagogical principles rather than used as a shortcut for task completion [2]. In Indonesia, the national higher-education guide on GenAI similarly recommends responsible integration, verification of outputs, and preservation of students' independent reasoning processes [3].

Recent empirical evidence presents a mixed picture. A meta-analysis by Chen and Cheung reports an overall positive effect of GenAI on university learning outcomes, while Bai and Wang show that interaction quality and output quality can influence learning through perceived usefulness, motivation, and self-efficacy [4]. Conversely, Weeks et al. find that GenAI usage can be associated with lower examination performance when the technology substitutes for learning effort. This tension suggests that short-term academic improvement and durable competence should be examined separately [5].

The distinction is especially important for skill retention. Grades may improve when students receive faster explanations or task assistance, yet durable learning requires retrieval, reflection,

metacognitive monitoring, and repeated practice. Research on self-regulated learning and learning science shows that students retain knowledge more effectively when they actively manage learning strategies and engage in cognitively effortful practice. Accordingly, intensive GenAI usage may be beneficial for some students but less effective, or even counterproductive, for others. Studies in Indonesia reinforce the need for contextual analysis. Indonesian students report using ChatGPT and related tools for academic writing, idea development, and learning support, while also recognizing concerns related to dependency, information validity, plagiarism, and academic integrity [1, 3, 5]. These findings indicate that institutional policy, AI literacy, study habits, and patterns of tool use should be considered when evaluating educational outcomes.

Most prior studies use surveys, perception analysis, descriptive statistics, correlational models, or predictive machine learning. Such approaches are useful for identifying patterns, but they do not directly estimate whether differences in GenAI usage cause differences in academic performance or skill retention. Learning analytics research has increasingly called for methods that support intervention-oriented decision making rather than prediction alone [4, 6]. Causal Machine Learning offers an appropriate computational framework because it can estimate treatment effects while controlling for multiple confounders and can identify variation in effects across student groups [6]. This study estimates the causal impact of intensive GenAI usage on students' GPA Change and Skill Retention Score using Double Machine Learning and Causal Forest. The study addresses three objectives: (1) to estimate the average effect of high GenAI usage on GPA change, (2) to evaluate its effect on skill retention, and (3) to identify heterogeneous effects across student characteristics and academic disciplines. By integrating learning analytics with causal inference, the study contributes a data-driven approach for evaluating GenAI adoption in higher education.

2. RESEARCH METHOD

2.1 Generative AI in Higher Education and Computing Education

GenAI systems generate text, code, images, and other digital content in response to user prompts. In higher education, these systems are increasingly used as on-demand learning assistants for explanation, brainstorming, summarization, feedback, and problem solving. UNESCO and the OECD describe GenAI as a potentially valuable educational resource, but both emphasize that its contribution depends on pedagogical design, governance, and responsible use [7]. The implications are particularly relevant to computer science and related fields. GenAI can generate programming solutions, explain algorithms, propose test cases, and assist with debugging. Zastudil et al. report that students and instructors in computing education see opportunities for support and curriculum innovation, while also raising concerns about overreliance, trustworthiness, plagiarism, and assessment validity. Kasneci et al., Lo, and Tlili et al. similarly argue that large language models can support personalized assistance and interactive learning, provided that learners evaluate outputs critically [8].

2.2 Academic Performance, Skill Retention, and Self-Regulated Learning

Academic performance and skill retention capture related but distinct outcomes. GPA change reflects observable academic achievement over a semester, whereas skill retention concerns the ability to preserve, transfer, and apply knowledge after initial learning. A technology may improve task completion without strengthening durable competence. This distinction is essential when AI tools can provide immediate answers or polished outputs. Current evidence is not uniform. Chen and Cheung synthesize studies showing that GenAI can improve learning outcomes across several educational settings [9].

Bai and Wang find that learning benefits are influenced by interaction quality, perceived usefulness, motivation, and self-efficacy. In contrast, Weeks et al. report lower examination performance among GenAI users in their observational study, indicating that extensive usage may hinder learning when students depend on generated answers instead of engaging with the underlying material [10]. Self-regulated learning provides a useful explanation for these differences. Zimmerman defines self-regulated learners as students who plan, monitor, and evaluate their own learning processes. Pintrich emphasizes the role of goals, motivation, and metacognitive strategies. Learning-science research

further shows that retrieval practice, self-testing, distributed practice, and desirable difficulties improve long-term retention. GenAI is therefore more likely to strengthen learning when it supports explanation, reflection, practice, and error correction rather than replacing cognitive effort [11].

2.3 Indonesian Higher-Education Context

The Indonesian context adds practical considerations related to access, literacy, policy, and academic culture. The national guide for GenAI use in higher education recommends that institutions treat AI as a learning aid, require verification of generated information, and develop ethical and pedagogical safeguards [12]. These recommendations are relevant because student usage often occurs outside formal institutional systems.

Empirical studies in Indonesia show that students use GenAI for academic writing, idea organization, and learning support. Alimin and Hariyadi identify patterns of GenAI use in academic writing among students in West Kalimantan [13]. Dewantara and Dewi discuss opportunities for efficiency and personalization alongside risks to integrity and independent thinking. Nursalam et al. report that students in Maluku use ChatGPT to support scientific writing while recognizing ethical and plagiarism concerns. Earlier work by Mairisiska and Qadariah documents positive student perceptions of ChatGPT as a digital learning resource. Together, these studies justify examining both benefits and risks in a quantitative causal framework [14].

2.4 Learning Analytics and Causal Machine Learning

Learning Analytics and Educational Data Science use learner data to understand and improve educational processes. Siemens and Baker describe the relationship between learning analytics and educational data mining as a basis for evidence-informed decision making [15]. Ferguson notes that the field has expanded from descriptive reporting toward predictive and prescriptive applications [16]. Nevertheless, prediction alone cannot determine whether an educational technology causes an observed improvement.

Causal inference focuses on estimating the consequences of an intervention or exposure. Athey and Imbens argue that causal evidence is necessary for policy evaluation because decision makers need to distinguish effects from associations. In observational educational datasets, this distinction is challenging because GenAI usage may be related to prior GPA, study habits, motivation, institutional rules, anxiety, or other confounding factors [17].

Double Machine Learning (DML) addresses high-dimensional confounding by separating the estimation of nuisance components from the estimation of the treatment effect. Causal Forest and Generalized Random Forest methods extend the analysis by estimating heterogeneous treatment effects across individuals or subgroups. Meta-learners and doubly robust approaches provide additional foundations for flexible treatment-effect estimation. These methods are suitable for educational datasets containing nonlinear relationships and mixed student characteristics [18].

2.5 Learning Analytics and Causal Machine Learning

The literature identifies both opportunities and risks associated with GenAI, but three limitations remain. First, much of the evidence is based on perception, correlation, or prediction rather than causal estimation. Second, many studies focus on academic performance without separately evaluating skill retention. Third, average effects may conceal substantial differences across students and academic disciplines.

This study addresses these limitations by treating intensive GenAI usage as an observational treatment and estimating its adjusted effects on GPA Change and Skill Retention Score. Cross-fitted Double Machine Learning is used to estimate average treatment effects after controlling for academic, behavioral, institutional, and psychological covariates. Empirical overlap, a continuous-treatment sensitivity model, and a quartile-based robustness test are evaluated explicitly. Heterogeneous effects are explored using a residualized random-forest model and are interpreted descriptively rather than as confirmatory subgroup causal claims. This analytical positioning aligns the study with computer-science research on learning analytics, interpretable machine learning, and data-driven educational decision support.

3. RESULTS AND DISCUSSION

3.1 Research Design and Workflow

This study applies an observational quantitative design within a Causal Machine Learning framework. The workflow separates descriptive inspection, predictive diagnostics, average-effect estimation, empirical-overlap assessment, sensitivity analysis, and exploratory heterogeneous-effect analysis. The distinction between predictive and causal stages is important: predictive accuracy indicates whether the recorded variables contain useful signal, whereas causal interpretation depends on design assumptions and the adequacy of the measured covariates.

Figure 1 summarizes the workflow. The average effects are estimated using cross-fitted Double Machine Learning (DML). Heterogeneous effects are examined using a stabilized residualized random-forest model. The latter analysis is explicitly exploratory because subgroup estimates can be sensitive to residual confounding, treatment-model errors, and multiple comparisons.

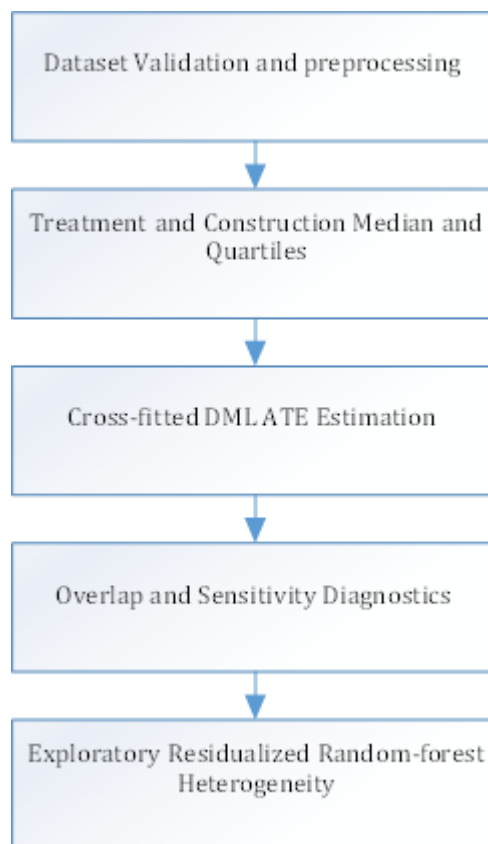


Figure 1. Analytical workflow for the causal machine learning analysis

3.2 Dataset Validation and Data Preparation

The analysis uses a secondary tabular dataset containing 50,000 student records and 16 original variables. The validation stage checks required columns, duplicate Student_ID values, and missing observations before model fitting. No duplicated identifiers and no missing values were detected. Student_ID is used only as a record identifier and is excluded from modeling.

Categorical variables are imputed using the most frequent value and represented through one-hot encoding. Numeric variables are imputed using the median and standardized. The dataset contains Burnout_Risk_Level, but this variable is excluded from the adjustment set and the primary subgroup analysis because its measurement time is unclear. Conditioning on a variable that may occur after GenAI usage could introduce post-treatment adjustment bias. GPA Change is calculated as follows:

$$GPA_{Change_i} = Post_{SemesterGPA_i} - Pre_{SemesterGPA_i} \quad (1)$$

This study uses a secondary dataset comprising 50,000 student records and 16 original variables. The dataset was obtained from [insert the actual data source, repository, or data-collection procedure]. Before model estimation, the data were examined for duplicate identifiers, missing observations, and consistency of the variables required for analysis. Student identifiers were excluded from the modeling process.

3.2.1 Measurement of Skill Retention

Skill Retention Score was used as the learning-outcome variable. Its operational definition, scoring range, measurement instrument, and assessment timing should be specified based on the original dataset documentation. In particular, the manuscript should clarify whether the score was obtained from a standardized assessment, a delayed post-test, a self-reported instrument, or another documented measurement procedure.

3.2.2 Adjustment Covariates

The adjustment set included baseline GPA, traditional study hours, major category, year of study, primary GenAI use case, prompt-engineering skill, tool diversity, paid subscription status, institutional policy, perceived AI dependency, and anxiety during examinations. These variables were included because they may be associated with both GenAI usage and the educational outcomes. Their temporal relationship to the treatment should be verified to reduce the risk of post-treatment adjustment bias.

3.3 Treatment, Outcomes, and Adjustment Covariates

The primary treatment is intensive GenAI usage. *Weekly_GenAI_Hours* is converted into a binary indicator using the sample median of 5.80 hours per week. Students at or above the threshold are categorized as high-usage students, while students below the threshold form the comparison group:

$$T_i = 1\{WeeklyGenAIHours_i \geq 5.80\} \quad (2)$$

The outcomes are GPA Change and Skill_Retention_Score. The adjustment set contains Pre_Semester_GPA, Traditional_Study_Hours, Major_Category, Year_of_Study, Primary_Use_Case, Prompt_Engineering_Skill, Tool Diversity, Paid Subscription, Institutional Policy, Perceived AI Dependency, and Anxiety_Level_During_Exams. These variables are treated as plausible pre-treatment covariates because they may be related to both GenAI usage and learning outcomes.

Table 1. Operationalization of the research variable

Variabel Group	Variabel	Analytical Role	Operational Definition
Treatment	Weekly GenAI Hours	Binary treatment	High usage if weekly hours ≥ 5.80
Outcomes	GPA Change	Academic outcome	Post-semester GPA minus pre-semester GPA
Outcomes	Skill Retention Score	Learning outcome	Recorded skill-retention score
Academic covariates	PreSemester GPA, Major Category, Year of Study	Adjustment set	Baseline academic characteristics
Behavioral covariates	Traditional Study Hours, Primary Use Case, Prompt Engineering Skill, Tool Diversity, Paid Subscription	Adjustment set	Study habits and GenAI-use characteristics
Institutional and psychological covariates	Institutional_Policy, Perceived_AI_Dependency, Anxiety_Level_During_Exams	Adjustment set	Contextual and psychological characteristics
Excluded from primary model	Student_ID, Burnout_Risk_Level	Not modeled	Identifier and

temporally ambiguous
variable

Because the data are observational, interpretation depends on consistency, empirical overlap, and conditional exchangeability. The analysis cannot eliminate bias from unobserved confounders. Therefore, the reported effects are interpreted as adjusted observational estimates rather than definitive proof of causality.

3.4 Cross-Fitted Double Machine Learning

DML separates nuisance-model estimation from treatment-effect estimation to reduce regularization bias. The partially linear model is written as:

$$Y_i = \theta T_i + \llbracket g(X) \rrbracket_i + \varepsilon_i \quad (3)$$

where Y_i is the outcome, T_i is the treatment, X_i is the confounding variable, $\llbracket g(X) \rrbracket_i$ is a nonlinear function learned by machine learning, and θ is a nonlinear nuisance function learned by machine learning, ε_i is an error term, and θ is the treatment-effect parameter. The treatment-assignment model is:

$$T_i = \llbracket m(X) \rrbracket_i + v_i \quad (4)$$

where $\llbracket m(X) \rrbracket_i$ estimates the conditional probability of intensive GenAI usage and v_i is the residual component. The implementation uses three-fold cross-fitting. For each held-out fold, Extremely Randomized Trees are trained on the remaining folds to estimate the outcome nuisance function and the treatment-assignment probability. The final estimate is obtained through residual-on-residual regression:

$$\hat{\theta} = \frac{\sum_{i=1}^n v_{1i} y_{2i}}{\sum_{i=1}^n v_i^2} \quad (5)$$

Heteroskedasticity-robust sandwich standard errors are used to construct 95% confidence intervals. The treatment model also produces out-of-fold propensity scores for empirical-overlap diagnostics.

3.5 Sensitivity and Exploratory Heterogeneity Analysis

Two sensitivity analyses are conducted. First, `Weekly_GenAI_Hours` is modeled continuously to estimate the adjusted marginal effect of one additional usage hour per week. Second, the upper and lower quartiles are compared directly. The quartile thresholds are 2.39 hours per week (Q25) and 11.72 hours per week (Q75). These analyses examine whether conclusions depend on the binary median split.

Exploratory heterogeneous effects are estimated using a residualized random-forest approximation. Individual pseudo-outcomes are constructed from the cross-fitted outcome and treatment residuals, stabilized by excluding very small treatment residuals and winsorizing extreme pseudo-outcomes at the 1st and 99th percentiles. The forest is fitted on a reproducible subsample of 15,000 records. The resulting conditional-effect estimates and subgroup means are descriptive and are not treated as confirmatory causal comparisons.

Predictive diagnostics are summarized using MAE, RMSE, and R-squared. Causal evaluation uses ATE values, heteroskedasticity-robust 95% confidence intervals, out-of-fold overlap diagnostics, continuous-treatment sensitivity, quartile robustness, and exploratory conditional-effect distributions.

$$ATE = Y_i(1) - Y_i(0) \quad (6)$$

3.6 Dataset Profile and Treatment Construction

The validated dataset contains 50,000 student records, no duplicated `Student_ID` values, and no missing observations. Figure 2(a) shows that GPA Change is concentrated around small positive values, whereas Figure 2(b) shows wider dispersion in Skill Retention Score. These distributions support the

use of both outcomes because short-term academic improvement does not necessarily imply durable skill retention.

Table 2. Treatment group division

Treatment Category	Criteria	Number of Students
High AI usage	Weekly_GenAI_Hours $\geq$ 5,8	25,008
Low AI usage	Weekly_GenAI_Hours $<$ 5,8	24,992

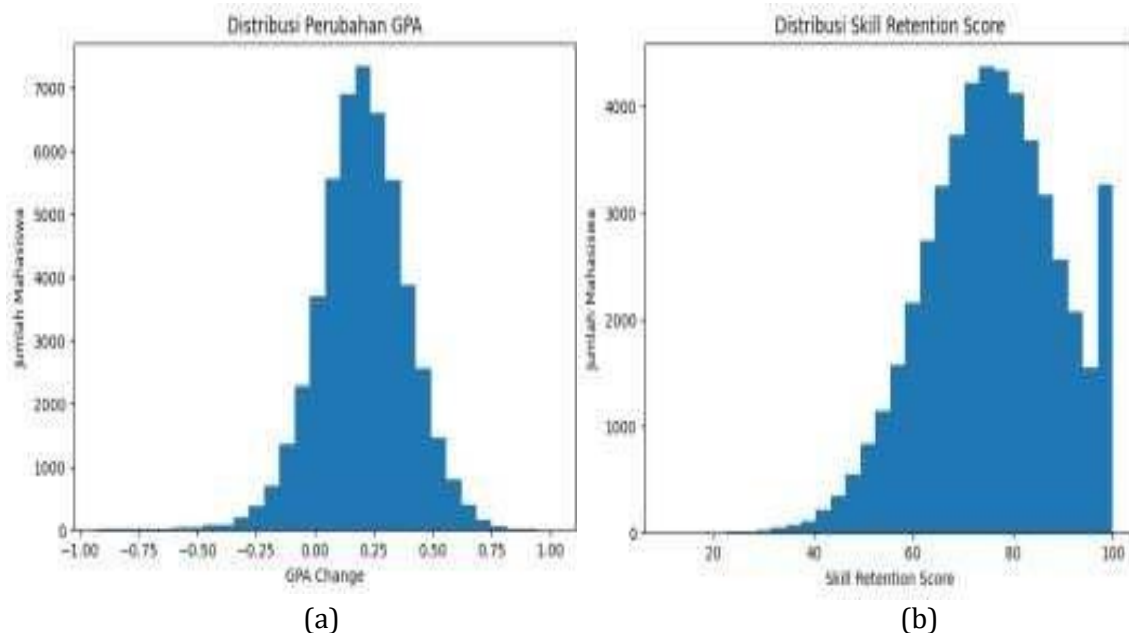


Figure 2. Distribution of changes in (a) student GPA and (b) student Skill Retention.

The median treatment threshold is 5.80 hours per week. This produces 25,008 students in the high-usage group and 24,992 students in the comparison group. The groups are almost balanced numerically, although balance in group size does not by itself guarantee causal comparability.

3.7 Dataset Profile and Treatment Construction

The holdout predictive model produced an MAE of 0.1229, an RMSE of 0.1558, and an R-squared value of 0.2933 for GPA Change. For Skill Retention Score, the corresponding values were 9.6624, 11.9720, and 0.1869. Skill retention is less fully explained by the recorded variables than GPA Change, which is consistent with the view that durable learning depends on processes not captured completely in a single tabular dataset. These values are predictive diagnostics, not causal evidence.

Table 3. Evaluation results

Outcome	MAE	RMSE	R ²
GPA Change	0.1229	0.1558	0.2933
Skill Retention Score	9.6624	11.9720	0.1869

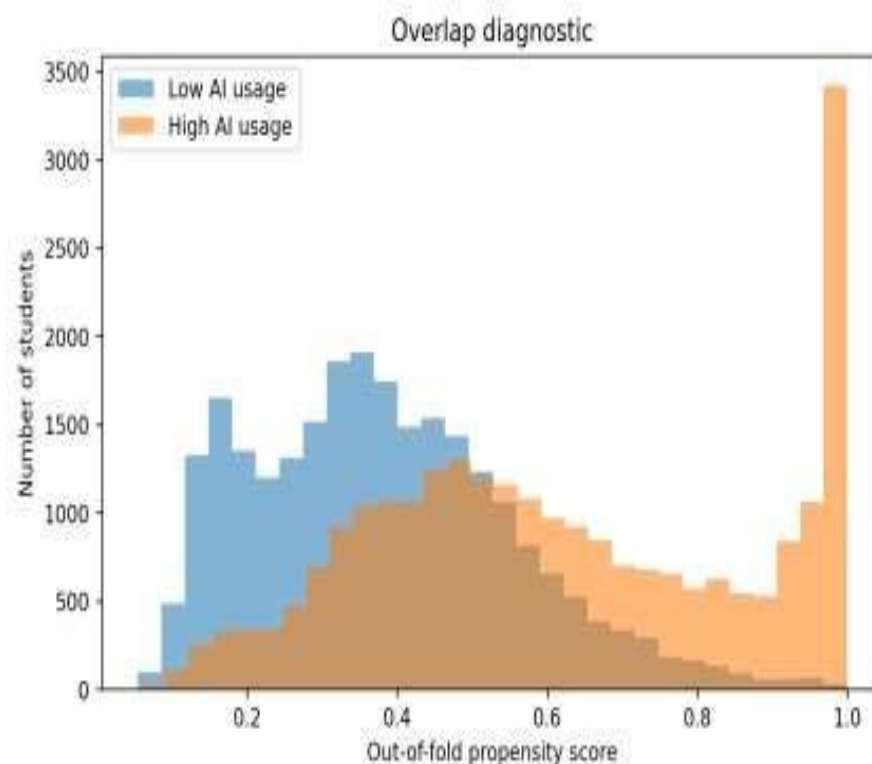


Figure 3. Empirical-overlap diagnostic based on out-of-fold propensity scores

Figure 3. presents the out-of-fold propensity distributions. The 1st and 99th percentiles are 0.101 and 0.999, respectively, while 8.56% of observations lie outside the [0.05, 0.95] range. Common support is broadly present, but the tail observations warrant caution. This diagnostic supports the plausibility of overlap without proving that all relevant confounders have been observed.

Table 4. Empirical-overlap diagnostic

Diagnostic	Value
Propensity-score p01	0.101
Propensity-score median	0.460
Propensity-score p99	0.999
Share outside [0.05, 0.95]	8.56%
OOF propensity ROC AUC	0.772

3.8 Main Double Machine Learning Estimates

The cross-fitted DML analysis estimates that high GenAI usage increases GPA Change by 0.0328 points on average, with a 95% confidence interval of 0.0297 to 0.0358. The estimated effect on Skill Retention Score is 0.8098 points, with a 95% confidence interval of 0.5733 to 1.0464. Both intervals exclude zero. After adjustment for the recorded covariates, intensive GenAI usage is therefore associated with better average outcomes under the binary treatment definition.

Table 5. Cross-fitted Double Machine Learning estimates

Outcome	ATE	SE	CI Lower	CI Upper	Interpretation
GPA Change	0.0328	0.0016	0.0297	0.0358	Positive; CI excludes zero
Skill Retention Score	0.8098	0.1207	0.5733	1.0464	Positive; CI excludes zero

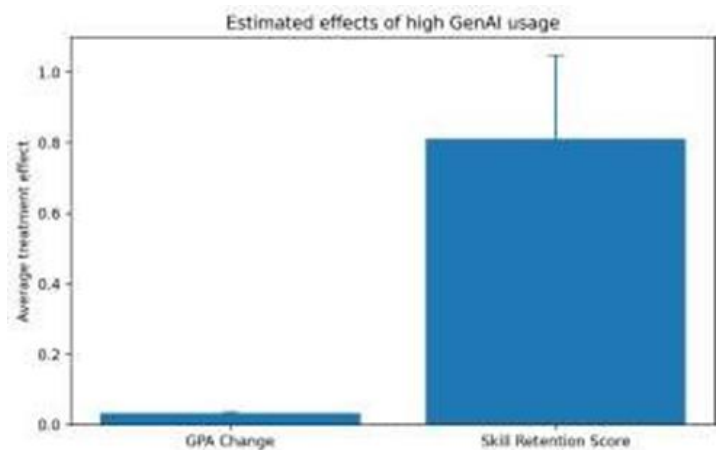


Figure 4. Estimated average effects of intensive GenAI usage on the outcomes.

These findings are consistent with prior studies showing that GenAI can support learning through rapid explanations, feedback, and adaptive support [8], [9]. However, the results should not be interpreted as unconditional evidence that more AI use always improves learning. The estimates remain dependent on observed covariates and the selected treatment definition.

3.9 Sensitivity Analysis

The continuous-treatment analysis estimates an average increase of 0.0007 GPA points per additional weekly GenAI hour (95% CI: 0.0005 to 0.0010). In contrast, the Skill Retention Score estimate is -0.1096 points per additional hour (95% CI: -0.1298 to -0.0895). The negative marginal estimate for skill retention differs from the positive binary-treatment estimate. This contrast suggests that the relationship is not a simple linear dose-response pattern: structured or moderate use may be beneficial, while additional usage intensity does not necessarily produce proportional long-term gains.

The quartile robustness analysis compares the highest-usage quartile with the lowest-usage quartile. The estimated effect is 0.0347 for GPA Change (95% CI: 0.0290 to 0.0403) and 0.8340 for Skill Retention Score (95% CI: 0.4093 to 1.2588). Both intervals exclude zero. The quartile contrast supports the direction of the binary-treatment findings, while the continuous specification cautions against assuming that each additional hour has the same educational value.

Table 6. Sensitivity analyses

Specification	Outcome	Estimate	CI Lower	CI Upper	Interpretation
Continuous: +1 weekly hour	GPA Change	0.0007	0.0005	0.0010	Positive marginal estimate
Continuous: +1 weekly hour	Skill Retention	-0.1096	-0.1298	-0.0895	Negative marginal estimate
Quartile: Q4 vs Q1	GPA Change	0.0347	0.0290	0.0403	Positive; CI excludes zero
Quartile: Q4 vs Q1	Skill Retention	0.8340	0.4093	1.2588	Positive; CI excludes zero

3.10 Exploratory Heterogeneous Effects

The exploratory residualized random-forest model indicates that the adjusted effect is not uniform across students. For GPA Change, the mean conditional estimate is 0.0350, with values ranging from -0.0456 to 0.0737. For Skill Retention Score, the mean estimate is 0.5344, with a range of -4.4046 to 5.9351. The distributions in Figure 5 show that some students may benefit more than others and that negative conditional estimates remain possible.

Table 7. Summary of exploratory heterogeneous-effect estimates

Outcome	Mean	STD	Min	Median	Max
GPA Change	0.0350	0.0178	-0.0456	0.0396	0.0737
Skill Retention Score	0.5344	1.6990	-4.4046	0.5766	5.9351

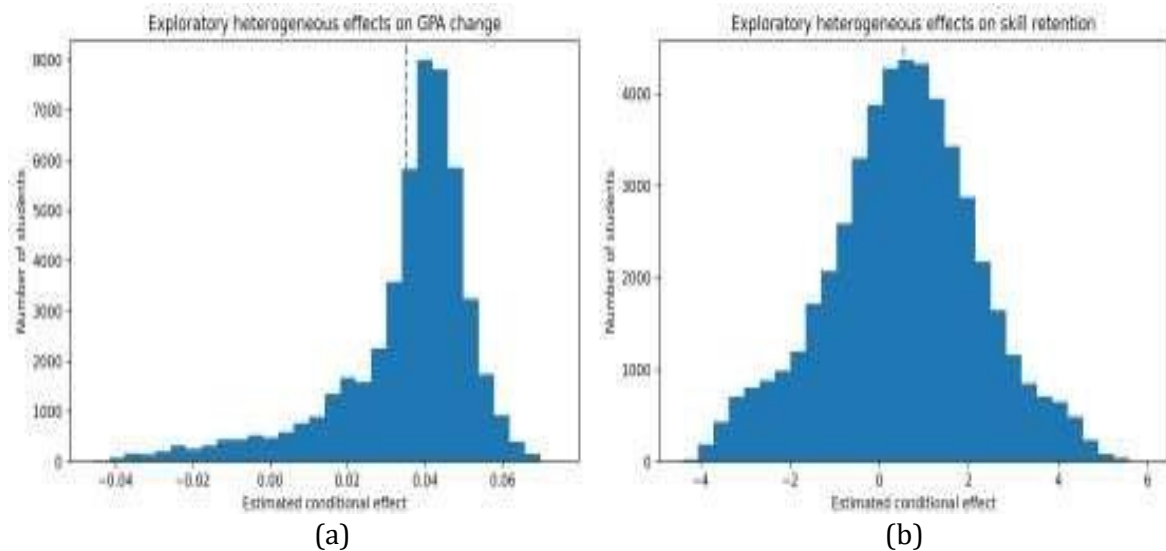


Figure 5. Exploratory conditional-effect distributions for (a) GPA Change and (b) Skill Retention Score

Figure 6 presents descriptive subgroup averages by major category. The estimated GPA effect is positive across all major categories and differs only modestly. The exploratory skill-retention estimates are higher for Humanities, Arts, and Business than for STEM. These contrasts may reflect differences in task structure and learning practices, but they should not be interpreted as confirmed causal differences between majors without targeted subgroup inference and stronger longitudinal measurement.

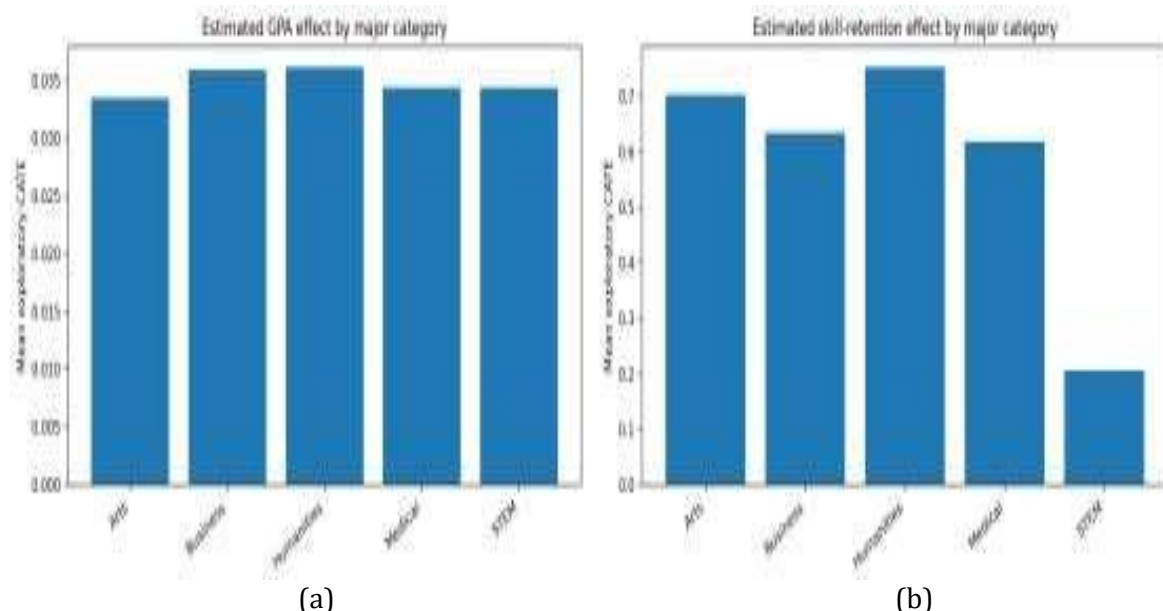


Figure 6. Exploratory mean conditional effects by major category for (a) GPA Change and (b) Skill Retention Score

The combined findings refine the interpretation of GenAI in higher education. GenAI can support academic performance and may assist skill retention when it is used as a learning aid. Nevertheless, the negative marginal skill-retention estimate in the continuous model indicates that additional usage intensity should not be equated automatically with deeper learning. This result is consistent with concerns that excessive reliance on GenAI may weaken independent cognitive engagement. Institutions should therefore prioritize AI literacy, process-based assessment, and assignments that require explanation, reflection, and independent problem-solving.

4. CONCLUSION

This study applies a cross-fitted Double Machine Learning workflow to evaluate the adjusted relationship between intensive GenAI usage and two educational outcomes: GPA Change and Skill Retention Score. Under the median-based binary treatment definition, high GenAI usage is associated with a higher GPA Change estimate (ATE = 0.0328; 95% CI: 0.0297 to 0.0358) and a higher Skill Retention Score estimate (ATE = 0.8098; 95% CI: 0.5733 to 1.0464) after adjustment for the recorded covariates. The comparison between the highest and lowest usage quartiles produces estimates in the same direction.

The sensitivity analysis provides an important qualification. When GenAI use is modeled as a continuous exposure, each additional weekly hour is associated with only a small positive GPA estimate but a negative marginal skill-retention estimate. The combined findings therefore do not support a simple assumption that more frequent GenAI use consistently produces deeper learning. The educational value of GenAI appears to depend on the purpose, quality, and context of use, including whether students employ the technology to understand concepts and practice reasoning or rely on it primarily to obtain answers.

The exploratory heterogeneous-effect analysis indicates that the estimated benefits differ across students and academic majors. These patterns are useful for generating hypotheses, but they should be interpreted cautiously because the dataset is observational, unmeasured confounding remains possible, and the timing of some variables requires stronger longitudinal documentation. Future studies should use longitudinal or experimental designs, define the temporal sequence of covariates explicitly, and validate the findings across institutions. For higher-education practice, GenAI should be guided as a structured thinking aid rather than a substitute for independent reasoning, retrieval practice, and sustained skill development.

REFERENCES

- [1] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.
- [2] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Education Sciences*, vol. 13, no. 4, Art. no. 410, 2023, doi: 10.3390/educsci13040410.
- [3] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel? ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, Art. no. 15, 2023, doi: 10.1186/s40561-023-00237-x.
- [4] C. Zastudil, M. Rogalska, C. Kapp, J. Vaughn, and S. MacNeil, "Generative AI in computing education: Perspectives of students and instructors," in *Proc. 2023 IEEE Frontiers in Education Conference (FIE)*, College Station, TX, USA, 2023, pp. 1–9, doi: 10.1109/FIE58773.2023.10343467.
- [5] UNESCO, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023.
- [6] OECD, *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. Paris, France: OECD Publishing, 2026, doi: 10.1787/062a7394-en.
- [7] S. S. Kusumawardani, D. Wulandari, P. Pannen, F. A. Ekadiyanto, I. M. Wiryana, A. Purwarianti, and S. A. I. Alfarozi, *Panduan Penggunaan Generative Artificial Intelligence pada Pembelajaran di Perguruan Tinggi*, ver. 1.1. Jakarta, Indonesia: Direktorat Pembelajaran dan Kemahasiswaan,

- 2024.
- [8] S. Chen and A. C. K. Cheung, "Effect of generative artificial intelligence on university students' learning outcomes: A systematic review and meta-analysis," *Educational Research Review*, vol. 49, Art. no. 100737, 2025, doi: 10.1016/j.edurev.2025.100737.
 - [9] Y. Bai and S. Wang, "Impact of generative AI interaction and output quality on university students' learning outcomes: A technology-mediated and motivation-driven approach," *Scientific Reports*, vol. 15, Art. no. 24054, 2025, doi: 10.1038/s41598-025-08697-6.
 - [10] J. O. Weeks, J. Voshaar, B. J. Plate, and J. Zimmermann, "Generative AI usage and exam performance," *arXiv preprint arXiv:2404.19699*, 2024, revised Nov. 2024.
 - [11] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educational Psychologist*, vol. 25, no. 1, pp. 3–17, 1990, doi: 10.1207/s15326985ep2501_2.
 - [12] P. R. Pintrich, "The role of goal orientation in self-regulated learning," in *Handbook of Self-Regulation*, M. Boekaerts, P. R. Pintrich, and M. Zeidner, Eds. San Diego, CA, USA: Academic Press, 2000, pp. 451–502.
 - [13] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology," *Psychological Science in the Public Interest*, vol. 14, no. 1, pp. 4–58, 2013, doi: 10.1177/1529100612453266.
 - [14] R. A. Bjork and E. L. Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, M. A. Gernsbacher, R. W. Pew, L. M. Hough, and J. R. Pomerantz, Eds. New York, NY, USA: Worth Publishers, 2011, pp. 56–64.
 - [15] H. L. Roediger III and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," *Psychological Science*, vol. 17, no. 3, pp. 249–255, 2006, doi: 10.1111/j.1467-9280.2006.01693.x.
 - [16] A. A. Alimin and H. Hariyadi, "Pola pemanfaatan generative AI dalam penulisan akademik mahasiswa Kalimantan Barat," *VOX EDUKASI: Jurnal Ilmiah Ilmu Pendidikan*, vol. 16, no. 2, pp. 455–466, 2025, doi: 10.31932/ve.v16i2.5516.
 - [17] B. A. Dewantara and L. K. Dewi, "Generative AI dalam pembelajaran mahasiswa: Antara inovasi pendidikan dan integritas akademik," *JiIP: Jurnal Ilmiah Ilmu Pendidikan*, vol. 8, no. 7, pp. 8209–8217, 2025, doi: 10.54371/jiip.v8i7.8680.
 - [18] Nursalam, J. Ambarita, and F. Resha, "Pemanfaatan ChatGPT dalam penulisan karya ilmiah mahasiswa perguruan tinggi di Maluku: Studi metode campuran," *Jurnal Pendidikan dan Kebudayaan*, vol. 10, no. 2, pp. 252–274, 2025, doi: 10.24832/jpnk.v10i2.6387.
 - [19] T. Mairisiska and N. Qadariah, "Persepsi mahasiswa FTIK IAIN Kerinci terhadap penggunaan ChatGPT untuk mendukung pembelajaran di era digital," *Jurnal Teknologi Pembelajaran Indonesia*, vol. 13, no. 2, pp. 107–124, 2023, doi: 10.23887/jurnal_tp.v13i2.2653.
 - [20] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, Vancouver, BC, Canada, 2012, pp. 252–254, doi: 10.1145/2330601.2330661.
 - [21] R. Ferguson, "Learning analytics: Drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, nos. 5–6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.