

Modelling a dataset-defined skill-retention outcome in generative AI-assisted learning: An exploratory machine-learning and explainability study

Clinton Amponsah ^{a *} , Bernard Kyiewu ^a  & Linda Bessa-Simmons ^a 

^a Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Ghana.

Suggested citation: Amponsah, C., Kyiewu, B., & Bessa-Simmons, L. (2026). Modelling a dataset-defined skill-retention outcome in generative AI-assisted learning: An exploratory machine-learning and explainability study. *Journal of Educational Technology and Online Learning*, 9(3), 568-581.

Highlights

- This exploratory study benchmarks prediction of a dataset-defined Skill Retention Score using a public 50,000-record synthetic/engineered GenAI-student dataset.
- The 0-100 retention field lacks documented instrumentation, delayed-assessment timing, reliability, and validity evidence; it is therefore treated as a benchmark proxy rather than a validated measure of durable learning.
- Extra Trees achieved $R^2 = .185$, $RMSE = 11.983$, and $MAE = 9.696$ on the reported hold-out split; permutation importance is interpreted only as a model-internal signal, with no causal, directional, or population-level claims.

Article Info: Research Article

Keywords: *Generative AI; skill-retention proxy; synthetic data; machine learning; learning analytics*

Abstract

Generative artificial intelligence (GenAI) is increasingly used in higher education, but durable learning cannot be inferred from task completion alone. This study presents an exploratory machine-learning benchmark using a public dataset of 50,000 student-like records that is treated here as synthetic/engineered because its source does not document an empirical sampling frame, institution, country, recruitment process, response rate, or primary-study ethics procedures. The target field, Skill Retention Score, is defined in the source schema on a 0-100 scale as representing skills retained and applied after the semester; however, the source provides no assessment instrument, delayed-assessment interval, reliability estimate, or validity evidence. Accordingly, the variable is analysed as a dataset-defined proxy rather than a validated measure of long-term retention. Ridge Regression, Decision Tree Regression, and Extra Trees Regression were compared in the originally reported 80:20 hold-out analysis. On that split, Extra Trees produced $R^2 = .185$, $RMSE = 11.983$, and $MAE = 9.696$. Its $RMSE$ is approximately 9.8% lower than the full-sample outcome standard deviation of 13.282, indicating modest predictive gain. The explainability output is permutation importance aggregated to the parent-variable level; it is interpreted as a model-internal, non-directional diagnostic, and SHAP results were not available in the submitted analytical record. The results therefore describe the structure of this benchmark dataset and the behaviour of the modelling pipeline; they do not establish effects of GenAI on real students, directional relationships, or instructional recommendations.

1. Introduction

Generative artificial intelligence (GenAI) is reshaping higher education by changing how students search for information, practise academic tasks, receive feedback, and regulate their learning. Unlike earlier educational technologies that mainly delivered content or automated grading, GenAI systems can produce explanations, examples, summaries, code, arguments, and personalised responses in real time. This creates new possibilities for learning support, especially where students require immediate clarification, alternative explanations, or guided practice. However, it also raises a deeper educational question: do students retain what they learn when GenAI assists the learning process? Recent studies show that GenAI is now widely used in higher education, yet its effects on learning remain mixed, context-sensitive, and pedagogically contested (Johnston et al., 2024; McGrath et al., 2025; Qian, 2025). Skill retention is a particularly

* Corresponding author. Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Ghana.
e-mail address: clinton.amponsah001@gmail.com

Doi: <http://doi.org/10.31681/jetol.1991829>

Received 5 Aug 2026; Accepted 22 Sep 2026

ISSN: 2618-6586. This is an open Access article under the CC BY license.



important but under-examined outcome in GenAI-assisted learning. While grades may reflect immediate task performance, skill retention concerns whether students preserve, recall, and apply knowledge after the learning event. This distinction is crucial because GenAI may help students complete tasks efficiently without necessarily strengthening durable competence. Weng et al. (2024) argue that GenAI challenges conventional assumptions about assessment and learning outcomes because educators must now distinguish between performance produced with AI assistance and learning genuinely retained by the student. Similarly, Pallant et al. (2025) found that GenAI supports higher-level learning when students use it to construct and augment knowledge, but weaker outcomes emerge when students use it procedurally. These findings suggest that GenAI's educational value depends on whether it promotes cognitive engagement rather than task completion alone.

The growing literature also indicates that the quality of GenAI engagement matters more than access or frequency of use. Students report using GenAI for writing, summarisation, idea generation, coding, and concept explanation, but they also express concerns about accuracy, academic integrity, reliability, and overdependence (Baidoo-Anu et al., 2024; Sousa & Cardoso, 2025; Kim et al., 2025). This matters for skill retention because students who critically evaluate, revise, and apply AI-generated outputs may retain more than students who passively accept generated answers. Prompt-engineering research reinforces this point. Lee and Palmer (2025) argue that prompt engineering should be treated as an academic competence because effective prompting can improve the educational value of GenAI interactions. Thus, the issue is not merely whether students use GenAI, but whether they use it in ways that support retrieval, reflection, transfer, and independent problem-solving. At the same time, GenAI-assisted learning introduces risks that may weaken retained competence. Heavy reliance on GenAI may reduce productive struggle, retrieval practice, and metacognitive monitoring, all of which are important for long-term learning. Abbas et al. (2024) found that GenAI use among university students has both helpful and harmful consequences, depending on learning motivation and usage behaviour. Zhang et al. (2024) showed that academic stress and performance expectations can contribute to problematic AI dependency, while Jia et al. (2025) found that GenAI dependency may negatively predict academic achievement through distorted self-efficacy. Academic-integrity reviews also warn that GenAI can blur the boundary between assistance and substitution, complicating how institutions evaluate authentic learning (Bittle & El-Gayar, 2025; Yusuf et al., 2024). These concerns make skill retention a stronger test of GenAI's educational value than short-term performance alone.

Despite these developments, existing research remains limited in two ways. First, much of the literature focuses on perceptions, academic integrity, adoption, or immediate performance rather than directly validated measures of retained learning. Second, explainable machine-learning studies can easily overstate what predictive associations mean when the underlying data source or outcome construct is incompletely documented. Accordingly, the present study is reframed as an exploratory methodological benchmark rather than an empirical study of a defined student population. It examines whether academic, behavioural, and GenAI-related fields in a public benchmark dataset contain predictive signal for the dataset-defined Skill Retention Score, while explicitly separating prediction from causation and model-internal importance from directional effects. The contribution is therefore methodological: to show what can and cannot be concluded from this dataset, to quantify the limited predictive gain of the models, and to identify the validation requirements that would be necessary before similar analyses could support claims about durable learning in real higher-education settings.

2. Literature

The literature on Generative AI-assisted learning is moving beyond adoption and academic-integrity debates toward deeper questions about retained competence, learning quality, and prediction. Skill retention is particularly important because students may complete tasks successfully with GenAI support without necessarily preserving knowledge for later use. Recent studies therefore suggest that GenAI should be examined through learning durability, feedback quality, student agency, and explainable analytics. This section reviews related studies under four themes that directly support the present study on predicting skill retention in GenAI-assisted learning.

2.1. Generative AI and Durable Learning Outcomes

Recent studies show that GenAI can support learning outcomes, but its contribution to durable learning depends on how students engage with it. Carmi (2025) found that students in higher education perceived GenAI as useful for academic learning, especially when it supported understanding, confidence, and learning efficiency. However, the study also implied that educational benefits are shaped by students' ability to use GenAI actively rather than passively. Pallant et al. (2025) provided stronger evidence that GenAI supports higher-level learning when students use it to construct and augment knowledge, rather than merely reproduce generated outputs. This distinction is central to skill retention because retained competence requires cognitive processing, revision, and transfer beyond immediate task completion. Weng et al. (2024) similarly argued that GenAI changes how learning outcomes should be assessed, since AI-supported performance may not always represent knowledge retained by the learner. These studies justify examining skill retention as a separate outcome from GPA or assignment completion.

2.2. Feedback, Assessment, and Retained Competence

Feedback quality is central to skill retention because students retain knowledge more effectively when they receive guidance that supports correction, reflection, and deeper understanding. Henderson et al. (2025) compared GenAI and teacher feedback and found that students evaluated feedback not only by usefulness but also by trustworthiness. This is important because AI-generated feedback may be immediate and accessible, but students still need judgement to decide whether it is accurate, relevant, and pedagogically meaningful. Haderlie et al. (2025) further showed that GenAI can influence student satisfaction and self-efficacy, suggesting that students' confidence in AI-supported learning may affect their engagement with academic tasks. Yet increased confidence does not automatically guarantee retained competence if students fail to verify, practise, or internalise what the tool provides. Weng et al. (2024) also emphasised that GenAI has implications for assessment validity because educators must distinguish assisted performance from authentic learning. These studies support the present focus on skill-retention prediction rather than short-term performance alone.

2.3. Prompt Engineering, Use Quality, and Learning Retention

A growing body of research suggests that the quality of student interaction with GenAI is more important than simple usage frequency. Lee and Palmer (2025) reviewed prompt engineering in higher education and argued that effective prompting can improve GenAI interactions by helping students obtain more relevant, accurate, and learning-oriented responses. Qian (2025) similarly identified prompt engineering as a key educational practice, especially where prompting strategies support critical skills and cognitive engagement. Park and Choo (2025) added that prompt writing, refinement, and optimisation are central to obtaining useful AI-generated outputs for educational purposes. These findings are highly relevant to skill retention because students who ask precise questions, compare outputs, request explanations, and revise prompts are more likely to remain cognitively active. In contrast, students who use GenAI mainly to obtain final answers may complete tasks without strengthening durable competence. Thus, recent studies support the inclusion of prompt-engineering skill and primary use case as predictors of skill-retention outcomes.

2.4. Machine Learning and Explainable Learning Analytics

Recent learning-analytics research supports the use of machine learning to model complex student outcomes, but it also stresses the need for transparency and interpretability. Khosravi et al. (2025) argued that GenAI is expanding the boundaries of learning analytics while requiring researchers to preserve validity, transparency, and pedagogical purpose. Misiejuk et al. (2025) mapped the emerging intersection of GenAI and learning analytics and showed growing interest in using AI to support learner monitoring, prediction, and educational insight generation. Rodríguez-Ortiz et al. (2025) reviewed machine learning and GenAI in higher education learning analytics, identifying predictive modelling and explainability as major directions for future research. Zhidkikh et al. (2024) also demonstrated the importance of generalisable and explainable predictive models in educational analytics. These studies justify the present comparative machine-learning approach. Since skill retention is multidimensional and difficult to observe

directly, explainable models can help identify which academic, behavioural, and GenAI-related variables provide meaningful predictive signals.

3. Methodology

3.1. Research Design

This study adopted a quantitative, exploratory predictive-modelling design. The dependent variable was the dataset-defined Skill Retention Score, a continuous field supplied in the public source dataset. The analysis was formulated as a supervised regression benchmark, not as a causal study and not as a population-inference study. The unit of analysis is therefore a student-like record in the benchmark file rather than a verified participant recruited through a documented higher-education sampling frame. The purpose was to examine how much predictive signal the available GenAI-use, academic, study-behaviour, and contextual fields contain for the target score and to evaluate the limitations of that prediction. Because the source is treated as synthetic/engineered benchmark data and the target construct lacks independent psychometric documentation, all substantive language about real students, durable competence, or instructional effects is avoided in the revised interpretation.

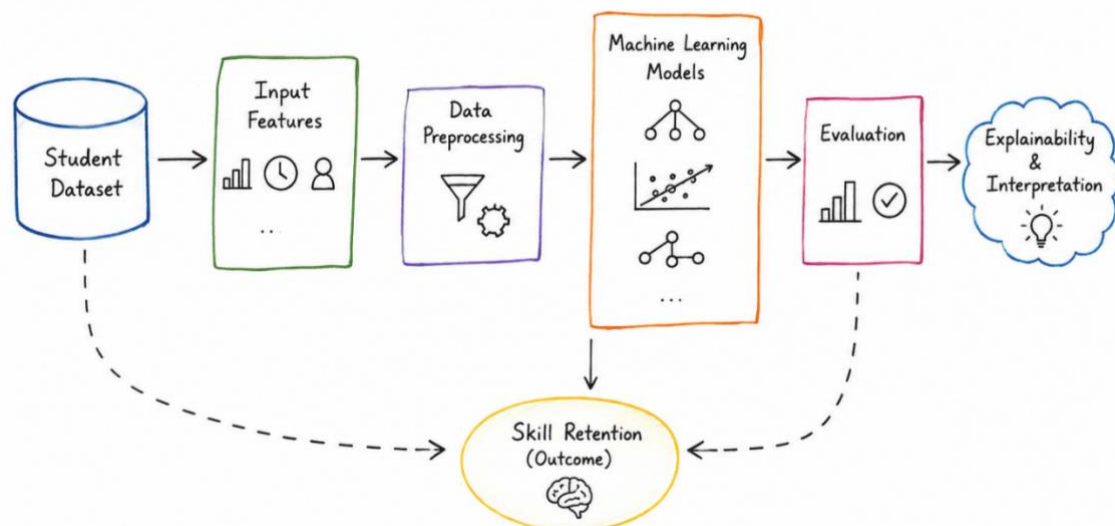


Fig. 1. Proposed architecture for predicting skill retention in Generative AI-assisted learning

3.2. Dataset and Study Variables

The analysis used the public “Impact of Ai on Students” dataset attributed to Nagi sisiro and distributed through Kaggle under a CC0 Public Domain licence (sisiro, 2026). The file contains 50,000 records and 16 source columns. The public source does not document an institution or institutions, country, data-collection period, recruitment or sampling procedure, response rate, or primary-study ethical approval and consent process. In addition, the highly structured public release is used widely as an educational/portfolio dataset. To avoid representing undocumented records as an empirical sample, the present revision treats the file as synthetic/engineered benchmark data. Consequently, the findings characterize the dataset and its apparent data-generating structure rather than a defined population of higher-education students. The predictor fields used in the retention model were pre-semester GPA, weekly GenAI hours, primary use case, prompt-engineering skill, tool diversity, paid-subscription status, traditional study hours, perceived AI dependency, institutional policy, exam anxiety, year of study, and major category. Student ID was excluded as an identifier, while post-semester GPA and burnout-risk level were excluded as co-outcomes to reduce leakage. The target, Skill Retention Score, is described in the source schema as a 0-100 score intended to represent how well skills were retained and could be applied after the semester. No item content, scoring rubric, assessor, assessment modality, interval between learning and testing, delayed-retention protocol, reliability coefficient, or construct/criterion validity evidence is supplied by the data source. The present study therefore uses the exact field as a dataset-defined proxy only. It cannot establish that the score is

psychometrically equivalent to delayed retention, durable competence, or an independently administered performance assessment.

3.3. Data Preparation and Leakage Prevention

Before modelling, the available fields were checked for the variables required by the analysis and column names were standardised for computational consistency. The target was separated from the predictor matrix, and Student ID, post-semester GPA, and burnout-risk level were removed from the predictors to avoid identifier use and co-outcome leakage. The preprocessing pipeline was configured to impute numerical variables with the median and categorical variables with the most frequent category if missing values were encountered. Numerical variables were standardised, and categorical variables were one-hot encoded. Preprocessing operations were fitted on the training partition only and then applied to the held-out partition. These steps prevent the held-out data from informing transformations fitted during model development; they do not, however, resolve the more fundamental limitations of source provenance and target construct validity described in Section 3.2.

3.4. Model Development

The numerical results retained in this revision reproduce the modelling design used in the submitted manuscript: one 80:20 train-test split followed by comparison of Ridge Regression, Decision Tree Regression, and Extra Trees Regression within a common preprocessing pipeline. Ridge Regression serves as a regularised linear comparator, while the tree-based models permit non-linear partitions and interactions. Importantly, the submitted analysis artefacts do not document a repeated nested cross-validation procedure, an inner-loop hyperparameter search space, selected tuned values, or a recoverable random-seed value. The revised manuscript therefore does not describe the models as tuned competitors and does not use the single split to establish a stable algorithm ranking. A confirmatory comparison would require model selection to occur entirely within the training data, preferably through repeated k-fold outer resampling with an inner tuning loop, followed by a genuinely untouched test set for the final estimate. Because those verified resampling outputs are not available in the submitted analytical record, no cross-validation means, confidence intervals, tuned hyperparameters, or test-set estimates are invented here; the reported results are explicitly limited to an exploratory hold-out benchmark.

3.5. Model Evaluation

Model performance was summarised with coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). R^2 describes the proportion of target variance accounted for on the reported split, RMSE expresses prediction error in target-score units with greater weight on larger errors, and MAE expresses the average absolute error. Ridge Regression provides a regularised linear benchmark. In addition, the full-sample standard deviation of the target (13.282) is used only as a descriptive scale reference for a mean-prediction benchmark. Because this standard deviation is not a separately computed held-out baseline, comparisons with it are presented as contextual rather than as a formal paired performance test.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

Root mean squared error was computed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

Mean absolute error was computed as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

where y_i denotes the observed dataset-defined Skill Retention Score for record i , $\hat{y}_i$ denotes its predicted value, $\bar{y}$ denotes the mean observed target value, and n denotes the number of observations evaluated. The originally reported metrics were calculated on the held-out 20% partition. They are presented as split-specific exploratory estimates. No statistical significance claim is made, and small numerical differences among algorithms are not interpreted as evidence of a stable ordering in the absence of repeated resampling and nested tuning.

3.6. Explainability Procedure

The explainability output reported in Table IV and Fig. 6 is permutation importance, as indicated by the original figure label “Mean Permutation Importance.” Permutation importance evaluates a fitted model by disrupting one predictor at a time and measuring the resulting deterioration in predictive performance. For categorical variables represented by multiple one-hot columns, the submitted analysis aggregated contributions back to the parent variable and normalised the construct-level values for presentation in Table IV. This clarification is important because the reported values are not the Extra Trees model’s default impurity-based `feature_importances_`. The reviewer’s concern about impurity importance favouring continuous or high-cardinality split variables therefore does not describe the metric plotted in Fig. 6. Nevertheless, permutation importance has its own limitations: correlated or redundant predictors can share or mask predictive information, construct-level aggregation depends on the encoding scheme, and the magnitude contains no sign or direction. SHAP outputs were not preserved in the submitted analytical record, so agreement between permutation importance and SHAP cannot be claimed. The revised paper therefore uses permutation importance only as a model-internal diagnostic and removes every directional, causal, or instructional inference from the ranking.

3.7. Reproducibility and Analytical Transparency

The analysis was implemented in Python using a pipeline that combined feature selection, preprocessing, regression modelling, evaluation, and generation of the reported outputs. The revised manuscript deliberately narrows its reproducibility claim. The public dataset is available from Kaggle as “Impact of Ai on Students” (<https://www.kaggle.com/datasets/laveshjadon/ai-impact-on-students>) under CC0. However, the submitted analytical record does not preserve the exact software/library versions, the random-seed value, a hyperparameter-search specification, or a deposited analysis-code archive. These omissions mean that the exact numerical results cannot be claimed to be fully reproducible from the manuscript alone. The reported metrics are therefore retained as the outputs of the original hold-out pipeline, while future confirmatory work should archive the complete environment, seeds, preprocessing specification, tuning grids, fitted parameters, and executable code alongside the data source.

3.8. Ethical and Methodological Considerations

This study is a secondary methodological analysis of a public, de-identified benchmark file and did not recruit participants, contact students, or collect new human-subject data. Because the public source does not document an empirical recruitment frame or primary-study ethics and consent procedures, this revision does not attribute such procedures to the dataset. Instead, the file is treated as synthetic/engineered benchmark data and the results are restricted to methodological and dataset-level interpretation. Student ID was excluded from prediction. The analysis is not intended for individual diagnosis, academic sanction, or automated educational decision-making. Any future study seeking conclusions about real students should use appropriately approved empirical data, informed-consent or other applicable governance procedures, a clearly operationalised delayed-retention assessment, and independent validation across institutions and learning contexts.

4. Results

4.1. Dataset-Defined Skill-Retention Outcome Profile

The outcome analysed here is the dataset-defined Skill Retention Score rather than an independently validated retention assessment. Student ID, post-semester GPA, Skill Retention Score itself, and burnout-

risk level were excluded from the predictor matrix as identifiers, the target, or co-outcomes. The regression models therefore used only the academic-background, GenAI-use, study-behaviour, and contextual predictor fields described in Section 3.2. Because the source is treated as synthetic/engineered, the following distributions describe records in the benchmark file and should not be interpreted as prevalence estimates for a real student population.

Table I: Descriptive Statistics of Key Variables for Skill-Retention Modelling

Variable	Mean	SD	Minimum	Median	Maximum
Skill Retention Score	75.798	13.282	10.780	76.000	100.000
Pre-Semester GPA	3.146	0.479	1.183	3.210	3.998
Post-Semester GPA	3.349	0.496	1.000	3.421	4.000
Weekly GenAI Hours	8.428	8.269	0.000	5.800	40.000
Tool Diversity	2.800	1.188	1.000	3.000	5.000
Traditional Study Hours	11.209	5.156	1.000	11.180	35.860
Perceived AI Dependency	3.505	1.821	1.000	3.000	10.000
Exam Anxiety Level	4.271	2.144	1.000	4.000	10.000

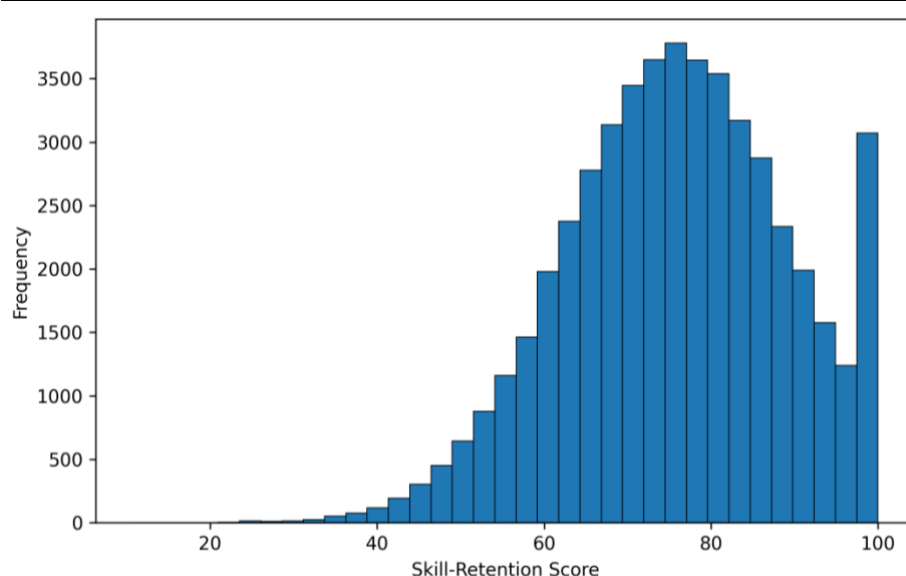


Fig. 2: Skill Retention Distribution

Table I shows that the dataset-defined Skill Retention Score had a mean of 75.798 and a standard deviation of 13.282, with a median of 76.000 and a range from 10.780 to 100.000. These values demonstrate substantial numerical dispersion in the benchmark target and therefore provide sufficient variation for an exploratory regression exercise. They do not, by themselves, establish meaningful variation in durable learning because the public source does not provide the instrument, assessment timing, reliability, or validity evidence required to interpret the field as a validated retention construct. Fig. 2 should therefore be read as the distribution of the supplied target variable only.

4.2. Descriptive Variation Across Generative AI Usage-Intensity Groups

For descriptive visualization, records were divided into four usage-intensity groups based on quartiles of weekly GenAI hours: Very Low, Low, Moderate, and High. The grouped variable was not entered as a separate predictor during model training. Because these groups are constructed from a predictor already present in the modelling dataset, they are not independent exposure conditions and the comparison cannot identify an effect of usage intensity.

Table II: Skill Retention by Generative AI Usage Intensity

GenAI Intensity Group	Weekly GenAI Hours Range	Mean Weekly GenAI Hours	Mean Skill Retention	SD
Very Low	0.00–2.39	1.132	75.653	12.876
Low	2.40–5.80	3.980	76.429	12.902
Moderate	5.81–11.72	8.410	77.117	12.809
High	11.73–40.00	20.204	73.995	14.280

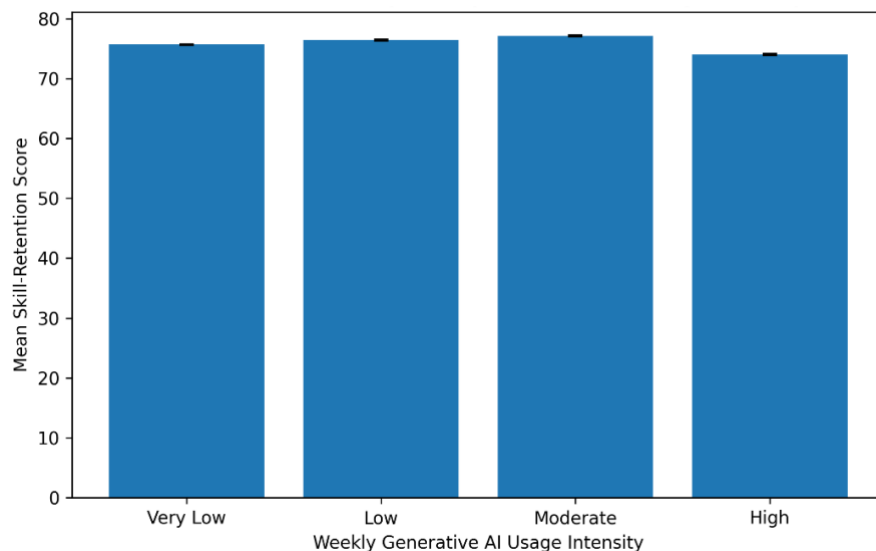
**Fig. 3:** Dataset-Defined Skill Retention Score by Weekly Generative AI Usage Intensity

Table II and Fig. 3 show a descriptive non-linear pattern in the benchmark data. The mean target score rises from 75.653 in the very-low group to 77.117 in the moderate group and is 73.995 in the high-usage group. No inferential test, effect size, confidence interval, or covariate-adjusted contrast was preserved in the submitted analysis for these grouped summaries, and plausible confounding by fields such as pre-semester GPA, traditional study hours, and major category is not addressed by the unadjusted means. Accordingly, the revised manuscript does not state that moderate use improves retention or that heavy use reduces it. The pattern is reported only as a descriptive property of the synthetic/engineered benchmark and is not used to support educational recommendations.

4.3. Exploratory Hold-Out Regression Performance

Ridge Regression, Decision Tree Regression, and Extra Trees Regression were evaluated on the originally reported held-out partition using R^2 , RMSE, and MAE. Because the analysis used a single split and does not include verified nested tuning or repeated cross-validation, the comparison is interpreted as an exploratory benchmark rather than as a definitive algorithm-selection exercise.

Table III: Exploratory Hold-Out Regression Performance for the Dataset-Defined Skill-Retention Score

Model	R^2	RMSE	MAE
Extra Trees	0.185	11.983	9.696
Ridge Regression	0.173	12.074	9.757
Decision Tree	0.123	12.433	9.997

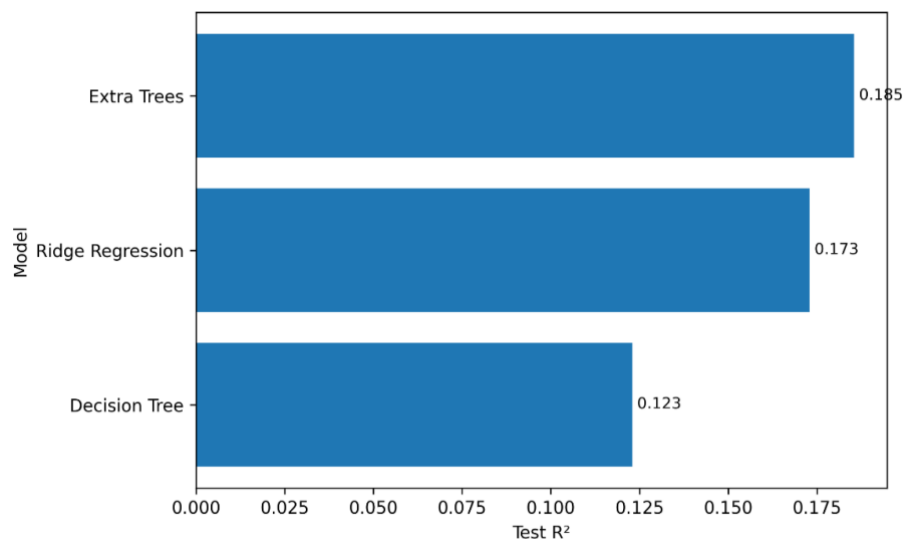


Fig. 4: Exploratory Hold-Out Regression Model Comparison

As shown in Table III and Fig. 4, Extra Trees produced $R^2 = .185$, RMSE = 11.983, and MAE = 9.696 on the reported split. Ridge Regression produced $R^2 = .173$, RMSE = 12.074, and MAE = 9.757, while the Decision Tree produced $R^2 = .123$, RMSE = 12.433, and MAE = 9.997. The numerical differences are small and are not treated as a stable ranking because variability across resamples was not estimated. For scale, the full-sample standard deviation of the target is 13.282; the Extra Trees RMSE is approximately 9.8% lower than that descriptive reference. This is a modest gain and does not warrant instructional prescription or high-stakes individual prediction. An R^2 of .185 is consistent with limited exploratory pattern detection in this benchmark. All comparisons with post-semester GPA have been removed because no equivalently specified GPA model was reported in the submitted analysis.

4.4. Agreement Between Predicted and Observed Skill-Retention Scores

The Extra Trees model, which had the lowest error on the reported hold-out split, was also examined using a predicted-versus-observed plot. The plot provides a visual check of how closely predictions from that fitted benchmark model align with the supplied target values; it is not an additional validation sample or a test of construct validity.

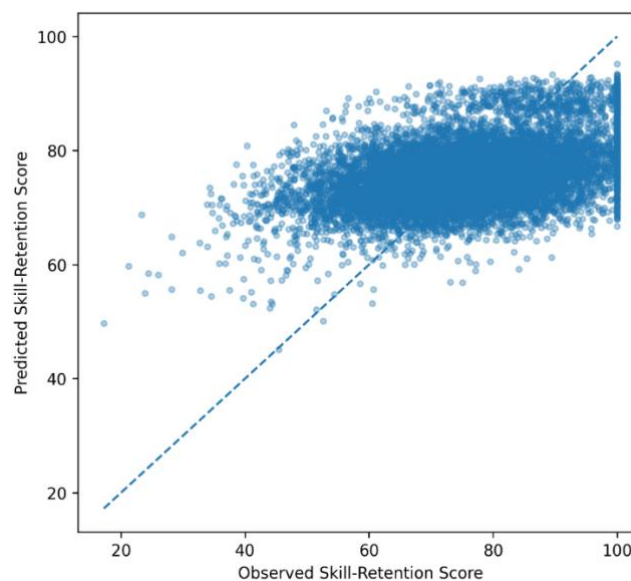


Fig. 5: Predicted vs Observed Skill Retention Score

Fig. 5 shows considerable dispersion around the diagonal reference line, consistent with the modest R^2 of .185. The model captures some broad variation in the supplied target but remains imprecise at the individual-record level. This visual result is therefore interpreted as evidence of limited predictive

resolution within the benchmark dataset, not as evidence that a particular learning process is inherently difficult to predict. The plot also cannot compensate for the absence of repeated resampling, external validation, or a validated delayed-retention measure. The appropriate conclusion is narrow: the available fields provide some predictive signal for this dataset-defined outcome, but the reported model is not suitable for high-stakes individual use and cannot support claims about real-world student retention without empirical validation.

4.5. Model-Internal Permutation Importance

Table IV reports construct-level permutation importance for the Extra Trees model shown in Fig. 6. The measure reflects the loss of predictive information when a variable is disrupted in the fitted benchmark model. One-hot encoded contributions were aggregated back to the parent-variable level for presentation. These values describe the dependence of this particular fitted model on the supplied predictors; they do not indicate whether higher values of a predictor increase or decrease the target, and they do not establish causal or educational importance.

Table IV: Aggregated Permutation Importance for the Reported Extra Trees Hold-Out Model

Rank	Predictor	Normalized Importance
1	Prompt Engineering Skill	0.280
2	Primary Use Case	0.127
3	Weekly GenAI Hours	0.088
4	Major Category	0.076
5	Year of Study	0.067
6	Traditional Study Hours	0.067
7	Perceived AI Dependency	0.058
8	Tool Diversity	0.056
9	Pre-Semester GPA	0.055
10	Institutional Policy	0.052
11	Exam Anxiety Level	0.043
12	Paid Subscription	0.030

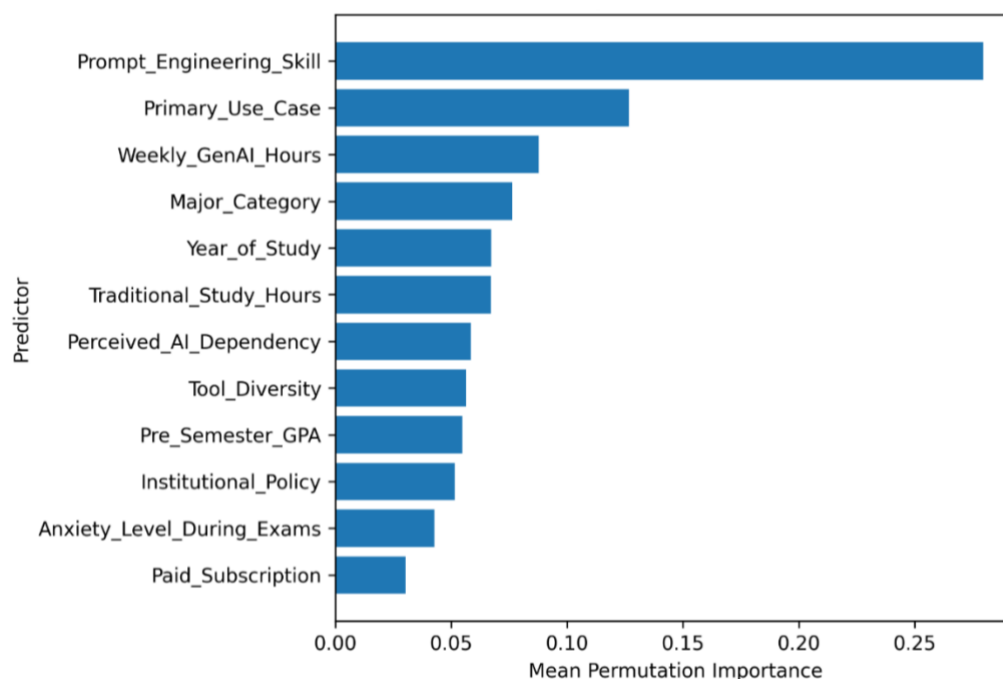


Fig. 6: Mean Permutation Importance for the Reported Extra Trees Hold-Out Model

Within the reported Extra Trees fit, Prompt Engineering Skill has the largest normalised aggregated permutation-importance value (.280), followed by Primary Use Case (.127) and Weekly GenAI Hours (.088). This ordering is a measure of model dependence in the benchmark, not an estimate of effect size or

direction. Permutation importance can be unstable when predictors are correlated or redundant because predictive information may be shared across variables, and aggregation of one-hot components makes the construct-level value dependent on the encoding and grouping rule. Because SHAP outputs were not preserved in the submitted analysis, the study cannot report whether this ordering is reproduced by SHAP. The revised manuscript therefore makes no claim that stronger prompt skill improves retention, that any use case is educationally superior, or that prior GPA is less important to real student learning. Table IV and Fig. 6 are retained solely as non-directional model diagnostics for the synthetic/engineered benchmark.

5. Discussion

5.1. *What the Benchmark Results Do and Do Not Show*

The revised results support a deliberately limited conclusion. The available predictor fields contain some signal for the dataset-defined Skill Retention Score, but the strongest reported hold-out model explains only 18.5% of its variance and reduces RMSE by roughly 9.8% relative to the full-sample outcome standard deviation used as a descriptive scale reference. These values are useful for judging the difficulty and potential utility of this particular benchmark task. They are not evidence that GenAI causes stronger or weaker learning, and they cannot be generalised to a campus population because the data source does not document empirical sampling or institutional provenance. The study's contribution is consequently methodological: it demonstrates how leakage prevention, baseline comparison, validation design, and restrained interpretation determine the credibility of GenAI learning-analytics claims.

5.2. *Construct Validity of the Dataset-Defined Retention Proxy*

The largest substantive limitation concerns the target itself. The source labels Skill Retention Score as a 0-100 measure of skills retained and applied after the semester, but it does not state what was assessed, how the score was constructed, who administered or scored it, when it was measured relative to the original learning episode, whether a delayed assessment was used, or whether the score has evidence of reliability or validity. The present analysis therefore cannot distinguish a true delayed-retention measure from a contemporaneous performance proxy generated or engineered for the dataset. This limitation changes the interpretation of every result: model accuracy is accuracy for the supplied field, not validated accuracy for durable human learning. Empirical research should operationalise retention prospectively with an explicit learning episode, a defensible delay interval, a documented scoring scale, and reliability and validity evidence before predictive results are linked to educational theory or practice.

5.3. *Limits of Model Comparison and Explainability*

The modelling and explainability outputs should also be read as preliminary. A single train-test split does not quantify performance variability and is insufficient for stable algorithm ranking when differences are small. A stronger design would use repeated outer cross-validation or a comparable resampling strategy, place all hyperparameter selection inside an inner training loop, report the search spaces and selected settings, and reserve an untouched test set for the final estimate. The submitted explainability figure already reports permutation importance, which is less directly exposed to the split-cardinality bias associated with impurity importance, but it remains non-directional and can be affected by correlated or redundant predictors. Robust follow-up analysis should compare permutation importance with SHAP and should use SHAP dependence, partial-dependence, accumulated-local-effects curves, or a clearly specified statistical model if directional interpretation is required. Until those analyses are verified, the current importance ordering is a diagnostic of one fitted benchmark model rather than evidence for educational mechanisms.

6. Conclusion

This study provides an exploratory machine-learning benchmark for a dataset-defined Skill Retention Score in a public 50,000-record GenAI-student dataset that is treated as synthetic/engineered. The source does not document an empirical sampling frame, and the 0-100 target lacks information about its assessment instrument, delayed-retention timing, reliability, and validity. On the originally reported hold-out split, Extra Trees produced $R^2 = .185$, $RMSE = 11.983$, and $MAE = 9.696$; the RMSE is approximately 9.8%

below the full-sample target standard deviation, indicating modest predictive gain. The reported permutation-importance values are model-internal and cannot establish direction or causation, and no SHAP agreement is claimed because SHAP outputs were not preserved in the submitted analysis. Accordingly, the study does not conclude that skill retention is harder to predict than GPA, that moderate GenAI use improves retention, or that prompt-engineering skill causes durable learning. The defensible contribution is a methodological one: the analysis illustrates both the predictive signal available in the benchmark and the limits that source provenance, construct validity, validation design, and explainability place on interpretation.

7. Future Work

Future work should begin with empirical, ethically governed data from clearly identified higher-education contexts and a prospectively defined retention outcome measured after a justified delay. The retention instrument should report its content, scale, scoring procedure, reliability, and validity evidence. Predictive modelling should use repeated nested cross-validation or an equivalent leakage-safe selection strategy, prespecify and report hyperparameter search spaces, archive seeds and software versions, and retain an untouched test set for final evaluation. Mean-only and unregularised/regularised linear baselines should be reported alongside flexible models. Explainability should triangulate permutation importance with SHAP, while directional claims should rely on SHAP dependence, partial-dependence or accumulated-local-effects analyses, or an explicitly fitted statistical model. Group comparisons of usage intensity should report sample sizes, confidence intervals, effect sizes, inferential tests, and appropriate adjustment for plausible confounders. Finally, code, environment files, and analysis outputs should be deposited with the manuscript so that all reported estimates can be independently reproduced.

Acknowledgments

The authors wish to express their sincere gratitude to the management and staff of the University of Energy and Natural Resources for providing the institutional support and resources necessary for this research. We also extend our appreciation to all students who participated in the study and contributed valuable data that made this analysis possible. Special thanks go to our colleagues in the Department of Computer and Electrical Engineering for their constructive feedback and encouragement throughout the research process. This study was conducted without external funding, and all authors contributed equally to the conceptualization, data collection, analysis, and writing of the manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT and Grammarly solely for language editing, grammar correction, and phrasing suggestions. No AI tools were used for data generation, statistical analysis, model development, or interpretation of results. The authors take full responsibility for all content.

References

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21, Article 10. <https://doi.org/10.1186/s41239-024-00444-7>
- Baidoo-Anu, D., Asamoah, D., Amoako, I., & Mahama, I. (2024). Exploring student perspectives on generative artificial intelligence in higher education learning. *Discover Education*, 3, Article 98. <https://doi.org/10.1007/s44217-024-00173-z>
- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), Article 296. <https://doi.org/10.3390/info16040296>
- Carmi, G. (2025). Learning with generative AI: An empirical study of students in higher education. *Education Sciences*, 15(12), Article 1696. <https://doi.org/10.3390/educsci15121696>

- Gao, Z., Cheah, J.-H., Lim, X.-J., & Luo, X. (2024). Enhancing academic performance of business students using generative AI: An interactive-constructive-active-passive (ICAP) self-determination perspective. *The International Journal of Management Education*, 22(2), Article 100958. <https://doi.org/10.1016/j.ijme.2024.100958>
- Haderlie, T. C., Jr., Sewak, S., Agarwal, S., & Lee, J. (2025). Investigating the effects of generative AI on student satisfaction and self-efficacy. *Marketing Education Review*. <https://doi.org/10.1080/10528008.2025.2576159>
- Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., & de Mello Heredia, J. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2025.2502582>
- Hon, K. K. L. (2026). Generative AI in higher education: A systematic review of its effects on learning outcomes and academic performance. *Journal of Educational Technology Systems*, 54(3), 537–560. <https://doi.org/10.1177/00472395251400089>
- Jia, W., Pan, L., & Neary, S. (2025). Effect of GenAI dependency on university students' academic achievement: The mediating role of self-efficacy and moderating role of perceived teacher caring. *Behavioral Sciences*, 15(10), Article 1348. <https://doi.org/10.3390/bs15101348>
- Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, 20, Article 2. <https://doi.org/10.1007/s40979-024-00149-4>
- Khosravi, H., Shibani, A., Jovanovic, J., Pardos, Z. A., & Yan, L. (2025). Generative AI and learning analytics: Pushing boundaries, preserving principles. *Journal of Learning Analytics*, 12(1), 1–11. <https://doi.org/10.18608/jla.2025.8961>
- Kim, J., Klopfer, M., Grohs, J. R., Eldardiry, H., Weichert, J., Cox, L. A., II, & Pike, D. (2025). Examining faculty and student perceptions of generative AI in university courses. *Innovative Higher Education*, 50, 1281–1313. <https://doi.org/10.1007/s10755-024-09774-w>
- Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22, Article 7. <https://doi.org/10.1186/s41239-025-00503-7>
- McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2025). Generative AI chatbots in higher education: A review of an emerging research area. *Higher Education*, 89, 1533–1549. <https://doi.org/10.1007/s10734-024-01288-w>
- Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics*, 12(1), 12–31. <https://doi.org/10.18608/jla.2025.8591>
- Pallant, J. L., Blijlevens, J., Campbell, A., & Jopp, R. (2025). Mastering knowledge: The impact of generative AI on student learning outcomes. *Studies in Higher Education*, 1–22. <https://doi.org/10.1080/03075079.2025.2487570>
- Park, J., & Choo, S. (2025). Generative AI prompt engineering for educators: Practical strategies. *Journal of Special Education Technology*, 40(3). <https://doi.org/10.1177/01626434241298954>
- Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends*, 69, 1105–1120. <https://doi.org/10.1007/s11528-025-01100-1>
- Qian, Y. (2025). Prompt engineering in education: A systematic review of empirical studies. *Journal of Educational Computing Research*. <https://doi.org/10.1177/07356331251365189>

- Rodríguez-Ortiz, M. Á., Santana-Mancilla, P. C., & Anido-Rifón, L. E. (2025). Machine learning and generative AI in learning analytics for higher education: A systematic review of models, trends, and challenges. *Applied Sciences*, 15(15), Article 8679. <https://doi.org/10.3390/app15158679>
- sisiro, N. (2026). Impact of Ai on Students [Data set]. Kaggle. <https://www.kaggle.com/datasets/laveshjadon/ai-impact-on-students>
- Sousa, A. E., & Cardoso, P. (2025). Use of generative AI by higher education students. *Electronics*, 14(7), Article 1258. <https://doi.org/10.3390/electronics14071258>
- Weng, X., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review. *Australasian Journal of Educational Technology*, 40(4), 1–16.
- Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, 40(6), 37–55. <https://doi.org/10.14742/ajet.9540>
- Yusuf, A., Pervin, N., Román-González, M., & Noor, N. M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, 21, Article 21. <https://doi.org/10.1186/s41239-024-00453-6>
- Zhidkikh, D., Heilala, V., Van Petegem, C., Dawyndt, P., Järvinen, M., Viitanen, S., De Wever, B., Mesuere, B., Lappalainen, V., Kettunen, L., & Hämäläinen, R. (2024). Reproducing predictive learning analytics in CS1: Toward generalizable and explainable models for enhancing student retention. *Journal of Learning Analytics*, 11(1), 132–150. <https://doi.org/10.18608/jla.2024.7979>