Understanding AI Assistant Usage Patterns in Academic Settings

Issar Manknojiya
Computer Science
Western University
Austin TX, USA
imanknoj@uwo.ca

Josh Cini
Computer Science
Western University
Toronto ON, CAN
jcini3@uwo.ca

Zachary Goodman
Computer Science
Western University
Toronto ON, CAN
zgoodma3@uwo.ca

I. Introduction

The widespread adoption of AI assistants such as ChatGPT has transformed academic workflows, yet their actual impact on student productivity remains poorly understood. While AI tools are now deeply embedded in coursework, study habits, and research assistance, educators and researchers lack empirical insight into how different usage patterns contribute to positive or negative academic outcomes. Understanding which session characteristics—such as task type, engagement level, assistance intensity, or student discipline—predict successful outcomes is essential both for optimizing AI system design and for helping students develop effective usage strategies.

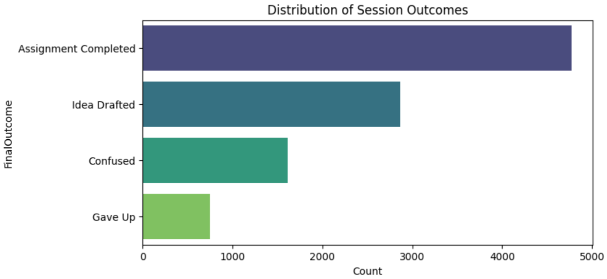
This study investigates these questions using a structured synthetic dataset replicating real-world AI usage patterns across diverse student groups. The dataset includes student classification variables (degree level and discipline), session characteristics (length, number of prompts, task type), engagement indicators (usage level, assistance level), and outcome metrics such as *FinalOutcome*, *SatisfactionRating*, and *UseAgain*. The primary goal is to model student productivity as a binary classification problem and identify the most influential behavioral and contextual factors driving successful AI-assisted study sessions.

Before applying machine learning methods, a detailed exploratory data analysis (EDA) is conducted to understand the distribution of student behaviors, inspect dataset balance, and observe potential relationships among features. These descriptive insights shape the modeling strategy and ensure interpretability of downstream results.

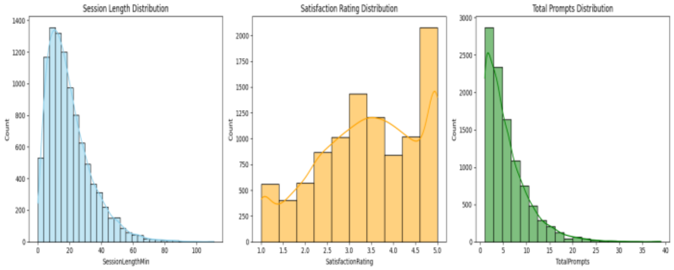
II. Exploratory Data Analysis

To contextualize the dataset and motivate the modelling approach, several descriptive analyses were performed. These visualizations highlight the dataset's structure, feature distributions, and target balance prior to training any machine learning models.

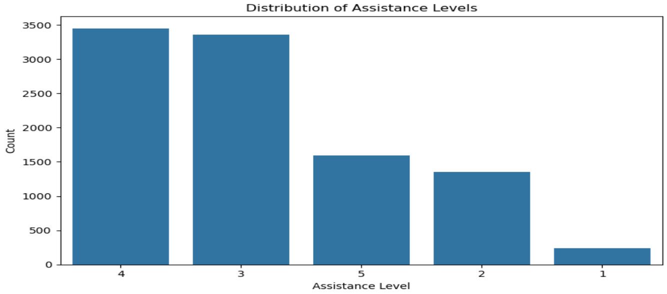
The *FinalOutcome* distribution provides an initial understanding of class balance by showing how many sessions resulted in “Assignment Completed” compared to unsuccessful attempts. This visualization is crucial because class imbalance affects learning algorithms and motivates the use of a stratified train–test split. Similarly, the *SatisfactionRating* histogram illustrates how students subjectively evaluated their AI-assisted sessions, revealing whether user sentiment tends to be positive, neutral, or polarized. These two variables form the core of the study’s productivity and experience assessment.

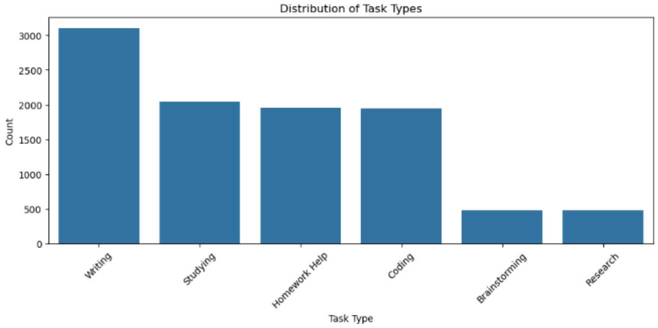


Histograms of *SessionLength* and *TotalPrompts* describe how students interacted with the AI tool during study sessions. The session length distribution highlights whether students typically use AI briefly or for extended work periods, while the prompt distribution reflects engagement level and conversational intensity. These behavioral patterns help explain variability in session outcomes and provide early clues about which features may correlate with productivity.



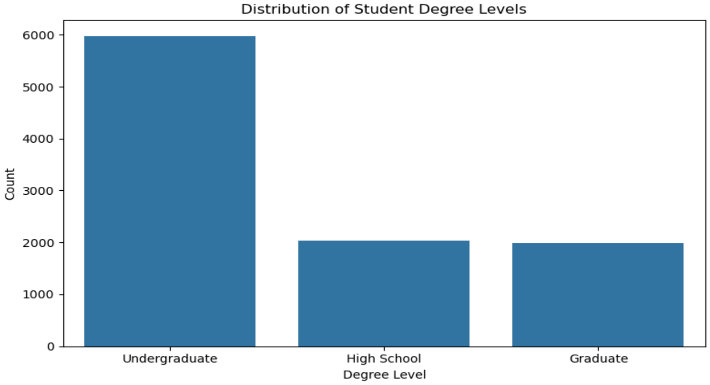
Visualizations of *AssistanceLevel* and *TaskType* illustrate how students rely on the AI system and what kinds of tasks they attempt to complete. The *TaskType* countplot shows the relative frequency of writing, coding, studying, math, and other tasks, which supports examining productivity differences across task categories. The *AssistanceLevel* distribution indicates whether students generally used AI lightly, moderately, or heavily—an important contextual factor identified in prior literature as influencing satisfaction and effectiveness.



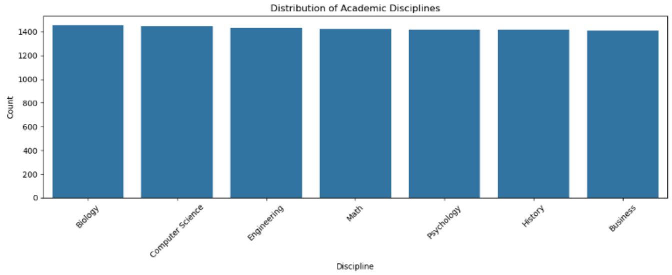


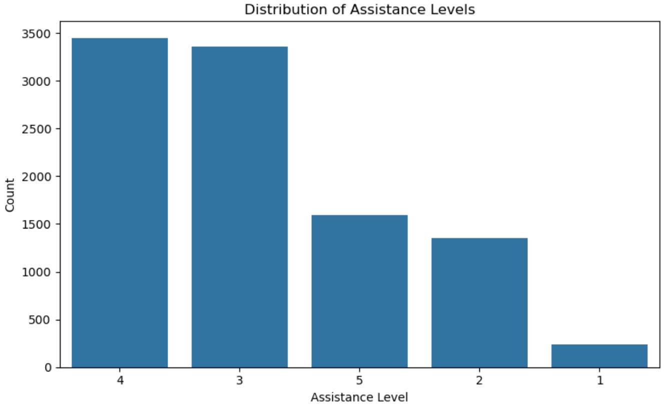
To understand how heavily students relied on the AI system during their sessions, a distribution of *AssistanceLevel* was examined. The countplot shows the relative frequency of low-, medium-, and high-assistance interactions. This distribution provides insight into the typical intensity of AI usage across the dataset and helps contextualize later modeling results. In particular, identifying whether most sessions involved minimal assistance or substantial involvement is important because assistance intensity is often linked to satisfaction, task complexity, and productivity outcomes in prior research. The visualization confirms a clear pattern of how students engaged with the AI tool, establishing an essential behavioral baseline before model training.

A countplot of *DegreeLevel* was generated to illustrate the representation of high school, undergraduate, and graduate students within the dataset. Understanding this distribution is critical because student academic level can influence both study strategies and AI usage patterns. For example, graduate students may undertake longer, more complex tasks, while high school students may use AI primarily for shorter or more guided interactions. This demographic breakdown ensures that any downstream analysis such as productivity modeling and considers potential differences in behavior across educational levels.



The distribution of academic *Discipline* was analyzed to determine how AI usage varies across fields such as STEM, humanities, business, and other academic areas. This visualization helps identify whether the dataset is balanced across disciplines or skewed toward certain fields. Since AI utility and interaction patterns can differ significantly by discipline—for example, coding tasks in STEM versus essay-writing tasks in the humanities, examining this distribution provides essential context for interpreting task-related trends and productivity outcomes. Including this demographic perspective strengthens the relevance of the modeling results by highlighting how diverse academic contexts shape AI-assisted learning behaviors.





III. Methodology

A. *Data Preprocessing*

Following the exploratory analysis, several preprocessing steps were applied to prepare the dataset for model training. First, all categorical variables. Including *TaskType*, *DegreeLevel*, *Discipline*, *UsageLevel*, and *AssistanceLevel*—were transformed using one-hot encoding to ensure compatibility with machine learning algorithms. Boolean variables such as *UseAgain* were converted to binary numerical form. Continuous variables (*SessionLength*, *TotalPrompts*, and *SatisfactionRating*) were scaled using a StandardScaler to normalize feature magnitudes and prevent any single variable from dominating the learning process.

The target variable, *FinalOutcome*, was converted into a binary label representing session productivity, where "Assignment Completed" was mapped to 1 and all other outcomes to 0. To maintain the original class distribution, the dataset was divided using an 80/20 stratified train–test split. This ensured that both training and testing sets preserved the balance between productive and unproductive sessions, reducing the risk of biased model evaluation.

B. *Feature Engineering*

Only standard transformations were used; no synthetic features were added to avoid inflating the influence of non-interpretable attributes. However, the preprocessing pipeline implicitly implements feature engineering through one-hot encoding, which expands categorical fields into high-dimensional binary vectors. This aids nonlinear models such as Random Forest by exposing category-specific interactions and allows linear models such as Logistic Regression to estimate per-category effects.

Additionally, continuous features (*SessionLength*, *TotalPrompts*, *SatisfactionRating*) were standardized to mean zero and unit variance. This step is essential for models with gradient-based optimization—particularly Logistic Regression—as unscaled features can cause poor convergence or unstable coefficient estimation.

C. *Model Selection and Hyperparameters*

Two supervised learning models were selected to provide both interpretability and performance-based comparison:

1) Logistic Regression (Baseline Model)

A Logistic Regression classifier was implemented using the lbfgs solver with an L2 regularization penalty and max_iter = 1000 to ensure convergence. The regularization strength parameter was set to C = 1.0, representing the default trade-off between fitting accuracy and coefficient shrinkage

Logistic Regression was chosen as a baseline due to its interpretability and ability to reveal linear relationships between session characteristics and productivity.

2) Random Forest Classifier (Nonlinear Model)

To capture potential nonlinear interactions not addressed by Logistic Regression, a Random Forest classifier was trained using 200 decision trees, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, bootstrap = True, and random_state = 42. This model is well-suited to heterogeneous, mixed-type datasets and provides robust predictive performance without requiring extensive hyperparameter tuning. The inclusion of Random Forest enables a comparison between simple linear models and more flexible ensemble methods.

D. *Evaluation Metrics*

Both models were evaluated on the held-out test set using multiple performance metrics: accuracy, precision, recall, F1-score, and ROC–AUC. Due to mild class imbalance observed in the target distribution, the F1-score and AUC were emphasized as the primary indicators of model quality. F1 captures the balance between false positives and false negatives, while AUC measures the classifier's ability to separate productive and unproductive sessions across all decision thresholds. This multi-metric evaluation provides a more reliable assessment than accuracy alone.

IV. Results

This section presents the performance evaluation of both Logistic Regression and Random Forest classifiers in predicting student productivity based on AI assistant usage patterns. Model performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics on the held-out test set. Additionally, feature importance analysis identifies the session characteristics most strongly associated with successful learning outcomes.

Both supervised learning models were trained on the preprocessed dataset and evaluated on a stratified test set containing 2,000 sessions (20% of the original dataset). The test set maintained the original class distribution, with 1,046 unproductive sessions (52.3%) and 954 productive sessions (47.7%). This mild class imbalance necessitated careful attention to metrics beyond simple accuracy.

1. Logistic Regression Results

The Logistic Regression baseline model achieved an overall test accuracy of 59.05%. The classification report revealed the following performance metrics:

Class 0 (Unproductive Sessions):

- Precision: 0.62
- Recall: 0.57
- F1-score: 0.59

- Support: 1,046 instances

Class 1 (Productive Sessions):

- Precision: 0.56
- Recall: 0.62
- F1-score: 0.59
- Support: 954 instances

Overall Metrics:

- Macro-averaged F1-score: 0.59
- Weighted-averaged F1-score: 0.59

The balanced F1-scores across both classes indicate that the model does not exhibit strong bias toward either productive or unproductive predictions. The slightly higher recall for productive sessions (0.62) compared to unproductive sessions (0.57) suggests that the model is marginally better at identifying sessions likely to result in assignment completion. However, the precision difference favors unproductive sessions (0.62 vs. 0.56), indicating that when the model predicts a session will fail, it is more often correct.

These results establish a reasonable baseline performance but leave substantial room for improvement, particularly given the potential nonlinear relationships in the feature space.

2. *Random Forest Results*

To capture potential nonlinear interactions between session characteristics, a Random Forest classifier was trained with hyperparameter optimization using 5-fold cross-validation grid search. The optimal hyperparameters identified were:

- n_estimators: 100 trees
- max_depth: 10 levels
- min_samples_split: 5
- min_samples_leaf: 2
- max_features: 'sqrt'

Using these parameters, the Random Forest model achieved:

- Test Set Accuracy: 59.65%
- Test Set ROC-AUC: 0.6303
- Best Cross-Validation Score (during training): 0.6113

The detailed classification report showed:

Class 0 (Unproductive Sessions):

- Precision: 0.62
- Recall: 0.57
- F1-score: 0.60
- Support: 1,046 instances

Class 1 (Productive Sessions):

- Precision: 0.57
- Recall: 0.62
- F1-score: 0.60
- Support: 954 instances

Overall Metrics:

- Macro-averaged F1-score: 0.60
- Weighted-averaged F1-score: 0.60

3. *Model Comparison and Interpretation*

The Random Forest classifier achieved a modest improvement of 0.6 percentage points in accuracy over Logistic Regression, with a corresponding increase in F1-score from 0.59 to 0.60. This marginal gain suggests that while some nonlinear feature interactions may exist, they do not dramatically alter predictive performance in this dataset.

The ROC-AUC score of 0.6303 for Random Forest indicates moderate discriminative ability—better than random guessing (0.50) but far from perfect separation (1.00). This score confirms that the model has learned meaningful patterns but that substantial overlap exists between the feature distributions of productive and unproductive sessions.

Both models demonstrate similar recall patterns: approximately 62% recall for productive sessions and 57% recall for unproductive sessions. This consistency suggests that the underlying feature space contains inherent ambiguity, regardless of model complexity. Several factors may explain the moderate performance:

a) Synthetic Data Constraints: The dataset is synthetic and may not fully capture the complexity of real-world student-AI interactions, including contextual factors such as prior knowledge, motivation, external distractions, and quality of AI responses.

b) Feature Overlap: The relatively balanced precision and recall values across both classes indicate that productive and unproductive sessions share similar session characteristics, making clean separation difficult.

c) Missing Contextual Variables: Important predictors such as student prior knowledge, learning style preferences, specific course requirements, and temporal factors (time of day, deadline proximity) are not captured in the dataset.

V. Future Improvements

One limitation of this study arises from the use of a fully simulated dataset, which—while useful for controlled experimentation—introduces randomness in certain attributes such as session dates, task distributions, and behavioral patterns. Because these values do not reflect true temporal, contextual, or academic pressures experienced by real students, some engineered features lack meaningful signal and may inadvertently reduce model clarity. As a result, model performance is constrained not only by the algorithms used but by the representativeness of the underlying data. Future work should incorporate more realistic or partially empirical datasets and apply domain-informed feature engineering (e.g., deadline proximity, cumulative workload, prior performance, or task complexity). These enhancements would allow models to capture deeper behavioral patterns and better distinguish productive from unproductive AI-assisted study session.