Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phase-Dependent Effects of Cognitive Autonomy on Student Performance in AI-Assisted Learning

Mohammed Alnahhal

mohammed.alnahhal@aurak.ac.ae

American University of Ras Al Khaimah

Nebiyu Gemedu

American University of Ras Al Khaimah

Adam Matar

American University of Ras Al Khaimah

Bayan Habis Alnaimat

Ministry of Education

Samariddin Makhmudov

Termez University of Economics and Service

Mosab I. Tabash

Al Ain University

Research Article

Keywords: Cognitive Autonomy Index, AI Tutoring, Self-Regulated Learning, Exam Performance, Dialogue Acts

Posted Date: July 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10021122/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial Intelligence (AI) tutoring in higher education has become an important part of the recent educational practices. This study introduces the Cognitive Autonomy Index (CAI), a theory-driven metric for characterizing student-AI interactions. The CAI captures the balance between behaviors that promote autonomy, such as asking conceptual questions, and those that undermine autonomy, such as requesting full solutions by AI tools. We analyzed longitudinal dialogue data from 110 undergraduates in an AI course to examine whether the CAI is associated with students' performance on examinations completed without AI assistance and whether these relationships vary across different phases of the course. Analyses reveal that aggregating autonomy-related behaviors into a single score obscures critical phase-specific dynamics. Overall AI usage volume showed only weak associations with examination performance. However, Principal Component Analysis (PCA)-derived behavioral components demonstrated stronger associations with performance, indicating that how students use AI tools matters more than how much they use them. Early-phase conceptual questioning was positively associated with examination performance, while late-phase context-dumping (pasting assignment material) was negatively associated with performance. Writing requests, classified as autonomy-undermining in CAI, was positively associated with early performance, revealing that the same behavior has different meanings at different times.

1. Introduction

Higher education started to incorporate heavily generative AI, particularly large language model chatbots. The percentage of full-time undergraduates in the UK who reported using AI increased from 66% in 2024 to 92% in 2025. Students reported using these tools to summarize information and explain concepts, but many are still unclear about what is considered appropriate use [1]. Some risks such as becoming overly dependent on AI and losing critical thinking abilities are revealed by some studies [2]. Students use GenAI for personal tasks, for academic work, and for a variety of assignments rather than just one [3]. These tools are now part of everyday student life.

Due to the rapid adoption, researchers have begun investigating how frequently AI is used. According to Cotton et al. [4] and Susnjak and McIntosh [5], universities had to respond swiftly to the threat of cheating. The frequency of using AI doesn't address the central question of this study: does GenAI assist students in thinking, or does it think for them? [3] In order to correct their own response, a student might, for instance, ask the AI to clarify a concept. Conversely, another student could copy the AI-generated solution in its entirety. Even though both of these activities qualify as "using AI," they differ greatly.

In this study, we evaluate the way students interact with GenAI tools. Therefore, we introduced cognitive autonomy index (CAI) to determine if students' prompts show that they are thinking or just handing the work to the AI tool. The definitions of cognitive autonomy support and self-regulation provide the base for our concept. Cognitive autonomy support provides reasoning and ownership for students. Self-regulation represents the planning, checking, and evaluation of one's own work over time [6, 7]. We use

two versions of CAI because a common pattern of behavior is potentially ambiguous: "providing context" may either represent providing a productive question to a GenAI tool or copying/pasting assignment materials so the tool completes them. CAI-A is conservative in treating this action, whereas CAI-B is stricter by weighting explanation-seeking more positively than direct production requests.

McNichols et al. [8] demonstrated that the manner in which students utilized an LLM tutor and the nature of interactions were more predictive of outcomes than the quantity of interactions. They categorized AI dialogue acts as either inquiry driven or solution driven. However, they did not develop nor test a method of measuring interaction quality related to cognitive autonomy, and they also did not investigate the effect at different times during a course. Our research builds upon McNichols et al. [8], in two manners. First, we measured the quality of AI interactions through CAI. This is a score for each student representing the proportion of autonomous cognition supporting behaviors compared to those that would replace autonomous cognition. Second, we investigated whether the effect of CAI scores changes over the semester. We analyze the effect of AI usage on early, mid, and late phases of the course.

Building on the StudyChat dataset for an undergraduate AI course, described by McNichols et al. [8] where data is licensed under CC BY 4.0., the current study employs a phase sensitive design: GenAI's associations with engagement differ across learning environments, and effects in the broader literature vary with implementation and duration rather than being uniform [9, 3]. The course was separated into three stages. In stage one, students were able to develop their foundational knowledge. Stage two involved practice and application of the skills developed during stage one. In the third stage, students applied all previously acquired skills while receiving increased assessments. Timing matters because some help can make the assignment look better without actually improving what the student knows. As such, only analyzing the final product would result in overestimating student learning relative to their actual knowledge [10, 9].

In order to demonstrate CAI using real student behavior, we used conversation logs from the StudyChat dataset collected within an undergraduate AI course and matched them to exam scores taken by students when no AI was utilized [8], and then determined whether CAI could accurately predict a student's performance on exams and if the ability to do so varied among the three distinct stages of the course (early, middle, and later). While the conversation logs provide valuable insight into student behavior post-AI interaction, they do not account for other actions performed by the student subsequent to interacting with the AI tool. Therefore, certain patterns observed in the logs may be due to characteristics of the AI system and not necessarily related to how the student utilizes or develops knowledge through the interaction [11].

2. Literature Review

The research regarding GenAI within the higher educational setting has been growing rapidly; however, much of this research still measures student usage solely by frequency or duration of usage [1, 3]. This review addresses three interrelated topics. First, adoption and integrity show why counting use came

first. The second topic concerns learning sciences that address why quality of use is relevant. Third, timing and conversation traces show why we need to test quality at different phases in the course [3, 9].

2.1. AI tutors and the habit of counting

The first point is about how many students use these tools. According to the Student Generative AI Survey, 92% of students in the UK have used GenAI, up from 66% the year before. Common applications include concept explanation and material summarization; 88% also report using it for assessments; many students are still unclear about what constitutes appropriate use [1]. Concerns about over-reliance, a lack of critical thinking, privacy risks, false information, and academic dishonesty are among the perceived benefits and concerns that are expressed in survey work conducted outside of the UK [2]. Similar findings from an extensive Chinese survey indicate that students apply these tools across a variety of task types and that academic GenAI use surpasses daily-life use, which already suggests that "use" is not a single, consistent behavior [3].

As such, educational institutions have been responding to these risks by addressing them in a couple of key areas. First and foremost, there were efforts made to protect assessment integrity. In addition to institutional pressures to maintain the integrity of assessments, this concern has impacted the early literature. Online examination research has shown that LLMs provide students with an advantage in producing sufficient answers in unproctored examinations. Therefore, universities have begun implementing stricter measures and redesigning their exams [5]. Additionally, research into academic integrity related to the use of AI-assisted writing has produced concerns regarding authorship and originality. Consequently, universities need to develop clear policies regarding the use of AI in assessments, increase staff training regarding its use, and implement a combined approach that includes both preventing cheating and detecting it when it occurs [12, 4]. When developing take-home assignments, educators should consider assessing student learning based on processes completed without the aid of AI. Broader policy work from UNESCO [13] treats these decisions as part of keeping humans in control of how AI is used in education [14].

Secondly, many educational institutions are considering GenAI as a tool to support learning goals. While some researchers believe that GenAI systems serve as tools to augment learner development and teacher decision making [15]. However, other scholars warn that educational institutions are adopting AI faster than they are designing good ways to use it, and we do not know enough about how it affects learning [15].

Despite these contributions, the literature shares a common gap: counting shows scale and urgency, but not whether students learn better. Two students who use AI can do very different things; one may be utilizing the system to enhance understanding and quality of work while the other is simply copying from the AI [3, 16]. Due to this limitation, the subsequent area of study focuses on student engagement, autonomy, and self-regulated learning.

2.2. Defining quality of use: engagement, autonomy, and self-regulation

A recent study published by Chi and Wylie [17] called the ICAP framework (interactive, constructive, active, and passive) provided a way for researchers to measure students' level of engagement as students interact with their learning environment. The study found that when students are able to construct upon their own ideas and build on those ideas rather than simply receive information, then they will be able to learn more effectively. Researchers focused on the area of autonomy support have studied areas such as choice, reasoning, and ownership. Researchers who focus on self-regulated learning have studied areas such as planning, monitoring and self-evaluation. When combined together, both of these frameworks provide a clear understanding of what good AI use looks like [7, 6] and they suggest that learning quality depends on active cognitive engagement, and not just receiving AI-generated answers.

Recent studies show clearly that better output does not mean better learning. An experiment was run by Fan et al. [10] using 117 university students. The results of the experiment showed that students who utilized ChatGPT for writing their essays produced higher quality essays but did not demonstrate greater knowledge transfer or greater actual learning. Their self-regulation patterns also changed. This supports previous warnings that AI may be able to produce a final product that appears to be of higher quality while providing little assistance to students understanding the material being presented [9]. Thus, the true question is not: will the output become better, but will the student learn.

Researchers have also explored how students utilize AI-generated output. Pallant et al. [18] distinguish a mastery approach, where students use GenAI to construct and augment knowledge, from a procedural approach, where students use it without knowledge augmentation, and they link the mastery pattern to higher-level learning outcomes [17]. Survey evidence also suggests that AI literacy and self-regulated learning relate to stronger writing performance and that writing performance partially mediates the link between AI literacy and GenAI-related well-being [19, 6]. What remains unresolved is whether those stronger outcomes depend more on the learner, the task, or the way the tool is guided. Recent studies also show that how students use AI depends on their learning behavior and skills. Rahman and Alam [20] found that self-directed learning affects how students benefit from AI tools. Other research also shows that self-regulated learning depends on both student characteristics and the features of AI tools. This means that students learn better when they actively manage and control how they use AI during their learning process [21]. Lim and Teh [22] showed that AI literacy and engagement help students take more control of their learning and improve outcomes.

The effect of AI use also depends on who the learner is and what the task requires. In student writing contexts, higher AI literacy and stronger self-regulation are associated with better writing performance [19, 9]. People who feel more confident using generative AI tools (like ChatGPT) also report that they spend less mental effort thinking critically. They choose to think less because they trust the AI. When AI

provides answers instantly, users may feel less need to check whether those answers are correct [23, 9]. Even if a student is motivated and smart, poorly designed assignments can still encourage lazy AI use.

Design studies suggest that guidance changes the story, but not automatically. In an experiment with undergraduates, guided LLM use produced stronger gains in writing quality and engagement than unguided use, with modest mental health benefits [24]. Debate-based instruction that integrated ChatGPT also reported gains in critical thinking and argumentation compared to traditional debate, though the context was course-specific [25, 9]. Frameworks like CHAT-ACTS explain how chatbots can support active learning, but without assessing what students achieve [26, 6].

Studies on using AI to provide feedback and to create AI-made educational materials have yielded both positive and negative results. Student's use of basic cognitive strategies has been shown to increase while their use of self-regulatory monitoring decreases as a result of receiving ChatGPT generated feedback [6, 27]. If students do not learn to effectively utilize this type of feedback, they will not be able to demonstrate improvements as a result of its use [6]. For example, research into AI created reading passages and teacher-created questions showed that the teacher-created question had significantly higher value in terms of promoting critical thinking than the AI-created passage [17, 28]. Researchers can currently identify theoretical descriptions of effective uses of AI in education. However, researchers continue to have difficulty identifying effective uses of AI in actual classroom settings.

2.3. Challenges in Interpreting Student–AI Interactions

The effectiveness of AI use by students may vary throughout the course of an academic term. A body of research referred to as the "Google Effect" indicates that when individuals expect that they will be able to obtain information easily, they will retain less information related to that topic and instead focus on locating it in the future [29]. The same concept may apply to student use of AI in tutoring sessions as answers are often available within one prompt. Additionally, large surveys also show that the connection between GenAI use and student engagement changes depending on the learning environment. Therefore, when and where students use AI determines whether the AI actually teaches them something or just does their work for them [9, 3]. Dialogue act research also explains why these conversation logs are hard to interpret. Model responses can show systematic dialogue-act patterns. These patterns often differ from how humans normally talk [11, 8]. The second constraint relates to scale and simulation. Simulation allows researchers to run many experiments quickly. However, simple or careless prompting can make student behavior look unrealistic. To get realistic results, researchers must make careful choices in how they set up the simulation [30, 11].

2.4. Gap statement

Although prior research has documented rapid GenAI adoption, a critical gap remains. Universities do not have a proven method for measuring how much students learn from their real conversations with AI tutors. They also do not know whether students who appear to learn well from AI actually perform better on tests and exams when they must work independently without AI assistance. The current paper addresses that gap. It develops and validates a way to assess the quality of real student–GenAI

interactions using autonomy-supportive and autonomy-undermining indicators. The paper also tests whether those interactions predict students' independent performance, and examines how these relationships vary throughout different stages of a course.

3. Research Questions and Hypotheses

This study addresses four research questions.

1. How well does the Cognitive Autonomy Index (CAI) predict performance on exams where AI use is prohibited?
2. Does CAI predict exam performance beyond simple usage volume measures, such as total interactions?
3. Do the CAI-to-exam-performance links remain stable after accounting for ambiguous dialogue behaviors, such as “provide context”?
4. Do links between specific AI tutor behaviors and exam performance change across early, mid, and late course phases?

Based on theories of self-regulated learning and prior evidence from educational technology research, we propose the following hypotheses:

- H1 (Usage Volume Baseline): Overall usage volume (e.g., total interactions) has a weak or nonexistent association with performance on AI-prohibited exams.
- H2 (Main Effect of Cognitive Autonomy): Higher CAI scores are positively associated with better performance on AI-prohibited exams after controlling usage volume.
- H3 (Phase-Dependent Effects): The relationships between particular AI tutor actions and exam performance vary over the early, middle, and late phases of the course.

4. Methodology

The methodology steps can be seen figure 1. It can be divided into three blocks (Data, Feature Engineering, and Modeling) and 8 steps. More details about its 8 steps are provided in the next subsections.

4.1. Data Sources, Course Context, and Integration

Two data sources were used in this study. The first one is the StudyChat interaction logs. McNichols et al. [8] developed the StudyChat Dataset, which is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The dataset contains student dialogues with ChatGPT collected within an artificial intelligence course environment. In total, there are 16,851 interactions from 203 students across Fall 2024 and Spring 2025. Students used a custom web application running gpt-4o-mini, the free ChatGPT tier, while working on seven programming assignments. The dataset is stored

in JSON Lines (.jsonl) format, where each line represents one student–AI interaction. Each record contains:

- userId: unique student identifier
- chatId: unique chat session identifier
- timestamp: interaction time
- prompt: student message
- response: AI tutor reply
- llm_label.label: hierarchical dialogue-act label, for example, conceptual_questions>Computer Science
- llm_label.justification: short rationale for the label

The second data source, which is about course assessment, is a CSV file with normalized scores per student. It includes AI-prohibited assessments (three exams or quizzes, e1, e2, e3, completed without AI support), AI-allowed assessments (seven assignments, a1 to a7, completed with AI support), and composite scores: no_llm (mean of e1 to e3) and llm (mean of a1 to a7).

The complete StudyChat dataset contains 203 students over two semesters [8]. We restricted the analytical sample to 110 undergraduate students who fulfilled two specific criteria: They agreed to share data from the tests, and their userId match in both the interaction logs (JSONL) and the assessment file (CSV). This match made it possible to track each student's use of AI and their performance in the course over time. The final sample of 110 students is 54 percent of the original StudyChat cohort. Each student met all inclusion criteria.

The course lasted 15 weeks and had three AI-prohibited tests. Exam 1 (e1) was about the first five weeks and assessed foundational concepts. Exam 2 (e2) was for weeks 6 to 10 and assessed intermediate applications. Exam 3 (e3) was about weeks 11 to 15 and evaluated integrative knowledge.

4.2. Variable Construction

4.2.1. Outcome Variable: AI-Prohibited Exam Performance

The primary outcome was no_llm, defined as the mean of the three AI-prohibited exams (no_llm = (e1 + e2 + e3) / 3), no_llm is continuous and goes from 0 to 1. This measure shows the performance of the student on the test without any help from AI. The source file made scores fit into the range of [0, 1]. Some students got extra points, which made the numbers go up (and sometimes greater than 1). For the β regression model, grades greater than 1 were truncated to 1. We shifted exact 0 and 1 values by epsilon = 0.001 (so all numbers are between 0 + 0.001 to 1- 0.001) to make it easier to estimate the boundary.

4.2.2. Predictor Variables

The predictor variables consisted of usage volume and behavior rates. Usage volume (total turns) was the total number of labeled AI interactions per student. For each student, dialogue-act category counts were converted into proportions by dividing each category count by the total number of turns. The analyzed categories included conceptual questions rate, representing the proportion of general knowledge questions; contextual questions rate, referring to questions about assignment-specific details or interpretation of outputs within the current context; writing request rate, indicating requests for generating text such as code, explanations, summaries, or conversions; provide context rate, representing interactions in which students supplied information to the LLM, including assignment instructions; verification rate, referring to requests for validation of code, writing, or outputs; editing request rate, indicating requests for modifications or improvements; off-topic rate, representing interactions unrelated to the assignment such as casual discussions, greetings, or expressions of gratitude; and miscellaneous rate, which included interactions that did not fit into any other dialogue-act category.

4.2.3. Cognitive Autonomy Index (CAI): Theoretical Grounding

To measure cognitive autonomy, we created the Cognitive Autonomy Index (CAI). The CAI shows whether students use AI to support their own thinking or not. Guided by self-regulated learning theory [6] and work on cognitive autonomy in classrooms [7], we classified each type of dialogue act as either autonomy-supportive (POS) or autonomy-undermining (NEG). We mapped each dialogue act to the StudyChat broad categories [8]. Autonomy-Positive Behaviors (POS) are conceptual questions (seeking explanations and "why" questions), contextual questions (problem clarification and parameter checking), verification (checking reasoning), and editing requests (revising self-generated content). For Autonomy-Negative Behaviors (NEG), two variants were defined to test robustness regarding the ambiguous "provide context" label. In CAI-A (Conservative), "writing request" is the only behavior that NEG includes. In CAI-B (Stricter), both "writing request" and "provide context" are part of NEG.

The CAI for each student was computed as:

$$CAI = pos_rate - neg_rate \quad (1)$$

Where,

$$pos_rate = POS / total\ turns \quad (2)$$

$$neg_rate = NEG / total\ turns \quad (3)$$

misc_rate and “off topic rate” are not included in both autonomy-positive behaviors and autonomy-negative behaviors.

However, CAI makes an assumption that the meaning of autonomy-supportive and autonomy-undermining behaviors is stable across time. As we show later, this assumption fails sometimes, where

behaviors classified as "autonomy-undermining" (e.g., writing requests) can support early learning, while "autonomy-supportive" behaviors may not uniformly predict success.

4.2.4. Dimensionality Reduction via Principal Component Analysis (PCA)

The seven behavior rates always add up to one for each student. If one rate goes up, others must go down. This creates a statistical problem called multicollinearity. If we include all rates in one regression model, multicollinearity inflates variance. We used Principal Component Analysis (PCA) to lower the number of dimensions. PCA transforms correlated predictors into a smaller set of uncorrelated components. Each component is a linear combination of the original variables.

Component retention was determined using three criteria: the scree plot (Fig. 2), eigenvalues greater than 1 (Kaiser criterion), and the interpretability of component loadings.
As will be shown in the results section, we kept PC1 and PC2. Together, they made for 45.5 percent of the total variance. PC1 accounted for 25.2%, and PC2 accounted for 20.2%. These parts had interpretable loading patterns. We used these orthogonal components as predictors in the regression models to reduce multicollinearity.

4.3. Data Preprocessing Pipeline

A number of steps were implemented in Python to construct the final dataset. At first, the hierarchical llm_label.label was parsed to extract the broad category (text before >). Then, rows with missing userId or label were removed. For each student, counts of each broad label were summed. After that, behavior rates and CAI variants were computed. Then, aggregated interaction data were merged with grades using userId. Finally, for beta regression, exam scores > 1 were truncated to 1, and exact 0/1 values were shifted slightly ($\varepsilon = 0.001$) to avoid estimation issues.

4.4. Analytical Strategy

The approach in this study contains several stages, beginning with descriptive and correlation analyses. We summarized the sample characteristics and the distribution of behavior rates and we calculated Pearson correlations between behavior rates and no_llm.

4.4.1. Overall Regression Models

We compared three models to test if AI interaction patterns predict exam performance better than usage volume alone.

Model 1: Baseline Model (OLS): It used ordinary least squares (OLS) regression to estimate the relationship between interaction volume and performance: *no_llm ~ total_turns*. This model captures the linear association between the number of StudyChat interactions and the average performance on AI-prohibited exams.

Model 2: PCA Model (OLS): It incorporated PCA to address multicollinearity among the behavior rates. The first two principal components (PC1 and PC2) were used as predictors together with interaction volume: *no_llm ~ PC1 + PC2 + total_turns*.

Model 3: PCA Model (Beta Regression): It retained the same predictors but employed beta regression instead of OLS: *no_llm_adj ~ PC1 + PC2 + total_turns*. Beta regression was used because exam performance is bounded between 0 and 1. The adjustment of the dependent variable (no_llm_adj) is a result of the truncated values and the shift of ε = 0.001 to satisfy the requirements of beta regression.

We utilized R^2 to assess the OLS models fit. However, pseudo-R^2 and AIC (Akaike Information Criterion) were used to assess beta regression. R^2 in OLS represents the proportion of variance explained by the predictors, whereas pseudo-R^2 in beta regression measures improvement over a null model. Lower AIC values indicate a better trade-off, which means explaining the data well without unnecessary predictors. AIC adds a penalty for each additional predictor.

4.4.2. Phase-Dependent Analysis

We developed a different beta regression model for each of the three exams: e1, e2, and e3. This is to examine the relationship between AI behaviors and test scores over time. Three factors are considered: theoretical relevance, the course structure at each phase, and early bivariate correlation patterns. In early phases, when students are focused on acquiring new concepts, conceptual questions are more important. By mid-course, during skill consolidation, too many contextual questions may give indication of confusion rather than real engagement. Late-phase (knowledge integration) is sensitive to any passive context-providing behaviors. We did not include additional behavior rates, like verification, editing, and off-topic requests because their proportions (compared to other types of behaviors) are usually low, and because they have weak importance.

The following phase-specific models were estimated:

- Early Phase: e1 ~ conceptual questions rate + writing request rate
- Mid Phase: e2 ~ contextual questions rate
- Late Phase: e3 ~ provide context rate

Effects were interpreted as log-odds (β) and converted to predicted probability changes for a 0.10 increase in behavior rate.

4.4.3. Robustness and Validity Checks

We conducted three robustness and validity assessments to ensure dependability and generalizability. The first one is ablation test. We checked to see if the composite CAI holds predictive information as well as its components. We compared a model with CAI to a broken-down model with no_llm ~ pos_rate + neg_rate + total_turns. If the decomposed model does not significantly enhance fit, this substantiates CAI as a concise and valid summary measure of autonomy-related AI activities. The second one is CAI

variant comparison. We tested sensitivity to how we classify ambiguous behaviors. We estimated all models employing both CAI variations. CAI-A considers writing requests as autonomy-undermining. CAI-B views writing requests and “provide context” as autonomy-undermining. Similar results across variants indicate robustness to this labeling choice. The third one is OLS vs. Beta regression comparison: We examined the coefficients and significance levels of OLS regression and beta regression because exam results are bounded to the range of 0 to 1. Beta regression is more appropriate for proportional outcomes, and consistency between the two models would strengthen the reliability of results derived from either technique.

4.4.4. Model Diagnostics and Assumption Checks

To ensure statistical robustness, we conducted comprehensive diagnostics. For multicollinearity, we calculated Variance Inflation Factors (VIFs) for raw behavior rates. If there are predictors with VIF values exceeding 5, then that indicates multicollinearity and, therefore, justifies the use of PCA. Test results are limited to a certain range (0–1), and educational scores often don't follow the normal distribution. Consequently, we employed beta regression as the principal method because it is appropriate for bounded proportional outcomes and does not require normally distributed residuals. Moreover, we used AIC values and pseudo-R^2 to select the best model specifications.

All analyses were conducted in R 4.4.3 using the following packages: tidyverse, betareg, car, corrplot, broom, kableExtra, and patchwork. Data preprocessing was performed in Python 3.9. Code is available upon request for reproducibility.

5. Results

5.1. Sample Characteristics and Descriptive Statistics

The final analytic sample was made up of 110 undergraduate students who had complete records of their interactions and assessments. This was 54% of the original StudyChat group. Table 1 shows the descriptive statistics for the most important variables.

Table 1
Descriptive Statistics of Key Variables (N = 110)

Variable	Mean	SD	Min	Max
AI Usage				
total_turns	88.26	89.14	1	386
Exam Performance				
no_llm	0.856	0.092	0.338	0.997
e1	0.819	0.145	0.403	0.99
e2	0.876	0.146	0	1.05
e3	0.873	0.114	0.57	1
CAI Variants				
CAI-A (Conservative)	0.129	0.431	-1	0.8
CAI-B (Stricter)	-0.033	0.479	-1	0.732

The average grades of the exams are above 0.81. The possible range is from 0 to 1. However, the maximum of 1.05 for the second exam (e2) was because of the bonus for some students. These numbers show that the distribution of students' grades cannot be normal. There were problems with the exam scores that needed to be fixed for beta regression. For example, e2 had 15 scores more than 1 (bonus points), 1 score equal to 0, and 1 score equal to 1. For e3, there were 2 scores equal to 1. Any score over 1 was truncated at 1, and any score that was exactly 0 or 1 was shifted by $\varepsilon = 0.001$.

Table 2 shows the distribution of AI behavior rates. Writing requests (34.0%) were the most prevalent activity, followed by conceptual questions (22.6%) and contextual questions (18.3%). The off topic and miscellaneous requests are rare. As mentioned before, both off topic and miscellaneous requests were not considered positive or negative in CAI.

Table 2
Mean Rates of AI Tutor Behavior Categories

Behavior	Mean Rate	SD
Writing requests	0.34	0.198
Conceptual questions	0.226	0.192
Contextual questions	0.183	0.154
Provide context	0.163	0.151
Verification	0.032	0.061
Editing requests	0.027	0.054
Off topic	0.027	0.069
Misc.	0.002	0.011

5.2. Bivariate Correlations with AI-Prohibited Exam Performance

Bivariate correlations (Pearson's r) between behavior rates and no_llm yielded preliminary insights (Table 3). There was a positive correlation between conceptual questions and test performance (r = 0.151, p = .116). However, there was a negative correlation between contextual questions (r = − 0.265, p = .005) and provide-context behaviors (r = − 0.209, p = .028) and exam performance. Both CAI variants demonstrated weak, non-significant negative associations. Given the weak bivariate correlations between both CAI variants and exam performance (r = -0.109 and − 0.047), we did not pursue simple regression models with CAI alone.

Table 3
Bivariate Correlations with AI-Prohibited Exam Performance (no_llm)

Variable	r	p-value
conceptual_questions_rate	0.151	0.116
writing_request_rate	0.187	.051†
editing_request_rate	0.15	0.119
verification_rate	−0.009	0.924
misc_rate	0.064	0.507
contextual_questions_rate	−0.265	.005*
provide_context_rate	−0.209	.028*
off_topic_rate	−0.171	.074†
CAI-A (Conservative)	−0.109	0.257
CAI-B (Stricter)	−0.047	0.624
total_turns	0.13	0.175
Note: † p< .10, * p < .05, ** p < .01.		

Figure 3 shows a heat map of bivariate correlations between AI tutor behaviors and exam performance. Warm colors, red and orange, show negative correlations. Cool colors, blue, show positive correlations. The strongest negative correlation is between conceptual questions and writing requests, r = -0.52. This means that students who focus on conceptual questions which means having behavior for promoting autonomy do not focus on writing questions which undermine autonomy.

5.3. Principal Component Analysis of AI Behavior Rates

To address the multicollinearity among behavior rates identified earlier, we conducted PCA. The first two components explained 45.5% of total variance. Table 4 shows the PCA loadings of the PC1 and PC2:

Table 4
PCA Loadings for the First Two Components

Behavior	PC1	PC2
Conceptual questions rate	0.424	0.586
Contextual questions rate	0.481	−0.124
Editing request rate	−0.119	−0.108
Off topic rate	0.201	−0.317
Provide context rate	−0.003	−0.616
Verification rate	0.144	−0.368
Writing request rate	−0.717	0.121

Misc_rate was excluded due to low frequency (0.2%). Off-topic rate was not central to our interpretation. The loadings reveal two distinct behavioral aspects. First, PC1 accounts for 25.2% of the variance. It contrasts high writing requests (negative loading: −0.72) with high conceptual questions (+ 0.42) and contextual questions (+ 0.48). This dimension reflects a contrast between writing request behavior and question-based interaction, which we label Cognitive Autonomy. Second, PC2 accounts for 20.2% of the variance. It contrasts high conceptual questions (positive loading: +0.59) with high “provide context” (negative loading: −0.62). This reflects a dimension ranging from deep conceptual engagement to more superficial contextual framing, which we label Problem Framing Quality.

Although PC3 had an eigenvalue higher than 1, PC3 was mostly loaded with editing requests, + 0.77, mixed with noise (off topic: −0.56). The other components have more fragmented loadings without clear patterns. Therefore, we selected only PC1 and PC2.

5.4. Overall Regression Models Predicting AI-Prohibited Exam Performance

Three models were compared to determine whether behavior patterns provide additional insight beyond overall AI usage (total turns) (Table 5).

Table 5
Comparison of Overall Regression Models

Model	R^2 / Pseudo-R^2	AIC	Key Predictors (β, p)
1. Baseline (OLS): no_llm ~ total_turns	0.017	−208.50	total_turns: 0.0001 (.175)
2. PCA-Based OLS: no_llm ~ PC1 + PC2 + total_turns	0.13	−217.93	PC1: −0.011 (.023)
			PC2: 0.019 (< .001)
3. PCA-Based Beta Regression: no_llm_adj ~ PC1 + PC2 + total_turns	0.14	−246.50	PC1: −0.110 (.026)
			PC2: 0.191 (< .001)
Note: Beta regression used boundary-adjusted scores (no_llm_adj).			

The PCA-based beta regression fit the data best. It produced pseudo-R^2 = 0.140 and the lowest AIC. This result shows that patterns of AI behavior explain more variance than usage volume alone, which supports H2. Behavioral research often treats pseudo-R^2 values between 0.10 and 0.20 as a meaningful improvement over a null model. However, it also means that there are other factors affecting student performance but were not included in this study.

5.5. Ablation Test: CAI Components vs. Composite Index

To test whether the composite CAI score works as a useful measure, we examined an ablation model: no_llm ~ pos_rate + neg_rate + total_turns.

This model's pseudo-R^2 of 0.138 is quite close to the PCA-based model's value of 0.140. Notably, pos_rate (β = 0.305, p = .012) and neg_rate (β = −0.218, p = .042) were both significant and in expected directions, even though the composite CAI score (pos_rate − neg_rate) showed weak, non-significant bivariate correlations with exam performance (Table 3). This suggests that while the individual components of CAI capture meaningful variance, aggregating them into a single score may obscure rather than clarify the relationship with learning outcomes.

5.6. Phase-Dependent Regression Results

Separate beta regression models were estimated for each exam phase (Table 6). Effects are expressed as log-odds (β) and as odds ratios (OR) for a 0.10 increase in behavior rate, where OR > 1 indicates higher odds of exam success and OR < 1 indicates lower odds. These are also converted to predicted probability changes for easier interpretation.

Table 6
Phase-Dependent Beta Regression Results

Phase	Exam	Predictor	β (SE)	p-value	OR (0.10 Δ)	Prob. Change (0.10 Δ)	Pseudo-R^2
Early	e1	Conceptual Questions	1.643 (0.407)	< .001***	1.18x	10.80%	0.147
		Writing Requests	1.206 (0.335)	< .001***	1.13x	8.00%	
Mid	e2	Contextual Questions	−1.152 (0.668)	.085†	0.89x	−10.6%	0.023
Late	e3	Provide Context	−1.662 (0.562)	.003**	0.85x	−15.9%	0.056

Note: SE = Standard Error of the beta coefficient (β), Prob. Change = approximate change in predicted exam probability; † p < .10, ** p < .01, *** p < .001.

5.6.1. Interpretation of Phase-Dependent Effects

For early phase (e1), conceptual questions showed a strong positive association. A 0.10 rise in the number of conceptual questions indicated a 10.8% rise in the odds of having better test performance. An OR of 1.18 means that a 0.10 increase in this rate is connected to an 18% higher chance of doing better on the test. Writing requests also showed a positive association. For mid Phase (e2), the initial model showed a negative association between contextual questions and exam performance (β = -1.152, p = .085). However, further investigation showed one influential case: student #57, who scored 0 on Exam 2 and had the highest contextual question rate. When this outlier was removed, the effect disappeared completely (β = 0.024, p = .972, pseudo-R^2 = 0.000). This indicates no reliable relationship between contextual questioning and mid-phase performance. For late Phase (e3), “provide context” behaviors were negatively associated with exam performance. A rise of 0.10 in this rate meant that the chance of passing the exam would go down by 15.9%. This pattern matches with using AI to complete assignments rather than learn from them.

These results partially support H3. The relationships between the AI interaction patterns and test performance change across phases. Early and late phases showed the predicted patterns: students who asked more conceptual questions generally performed better on the exam, whereas students who mainly provided context without further analysis tended to perform worse. However, the effect of contextual questions during the middle phase disappeared after removing outliers. Further investigation identified student #57 as an influential case in the mid-phase model. This observation illustrates how extreme values can affect model estimates and suggests that the mid-phase result does not reflect a general pattern.

5.6.2. Effect Size Interpretation

The magnitude of phase-dependent effects also warrants attention. Early conceptual questions exhibit a strong positive association with performance (β = 1.643), corresponding to an odds ratio of approximately 1.18 for each 0.10 increase in the rate of conceptual questions. In contrast, late-stage provide-context requests show a substantial negative relationship (β = −1.662), with an odds ratio of approximately 0.85 per 0.10 increase. Mid-phase contextual questions, however, do not exhibit a consistent effect across the population.

These effects are important for education, even though the pseudo-R^2 values are low. This is because they show that the chances of doing well on the exam change by 8–16% for every 10% change in behavior in the early and late phases. However, contextual questions in the mid-phase showed no consistent relationship with exam performance.

5.7. Robustness Checks

Table 7 shows that the direction of effects was uniform across both OLS and beta regression for every predictor. In other words, early conceptual and writing requests were positive in both, mid contextual and late “provide context” were negative in both. Statistical significance was stable for three out of four predictors. For mid-phase contextual questions, the negative effects were mainly influenced by a single outlier (student #57).

Table 7
Comparison of OLS and Beta Regression Coefficients

Phase	Predictor	β (OLS)	p (OLS)	β (Beta)	p (Beta)	Dir. Consist?	Sig. Consist?
Early	Conceptual	0.245	< .001	1.643	< .001	Yes	Yes
Early	Writing	0.177	0.001	1.206	< .001	Yes	Yes
Mid	Contextual	−0.249	0.004	−1.152	0.085	Yes	No
Late	Provide Context	−0.164	0.031	−1.662	0.003	Yes	Yes

*Note: The mid-phase significance inconsistency is explained by a single outlier (student #57).

Both CAI variants showed weak negative correlations with exam performance across phases (Table 8). CAI is a net score, not individual behaviors (CAI = pos_rate − neg_rate). If a student asks many conceptual questions (positive) but also makes many writing requests (negative), their CAI could be near zero or even negative. The correlation captures the net relationship, not the effect of individual behaviors. The contradiction comes from phase-specific dynamics. Behaviors considered autonomy-undermining in CAI, such as writing requests, showed benefits in the early phase. The same behaviors helped in some phases but not others. For example, writing requests were positively associated with early-phase performance but not later performance. Even "good" behaviors (conceptual questions) may not help equally in all phases. When aggregated into a single composite score, these opposing phase-

level patterns offset one another, giving the false impression that autonomy-related behaviors hurt performance. A single score like CAI can't capture the full picture. Understanding AI's impact means looking at which behaviors students use and when.

Table 8
Correlations of CAI Variants with Exam Performance by Phase

CAI Variant	e1	e2	e3	no_llm
CAI-A (Conservative)	−0.067	−0.127	−0.064	−0.109
CAI-B (Stricter)	−0.023	−0.078	−0.005	−0.047

Initial VIF results for the raw behavior rates showed significant multicollinearity. Two predictors went above the usual limit of 5. The VIF for writing request rate was 8.70, and for conceptual questions rate was 6.37. To solve the problem, PCA reduced multicollinearity to acceptable levels. All of the VIF values were less than 2: PC1 (1.06), PC2 (1.01), and total_turns (1.07). The decision to retain the initial two components was based on scree plot data (Fig. 2) and their interpretability. As mentioned before, we can say that PC1 stands for Cognitive Autonomy, and PC2 stands for Problem-Framing Quality.

6. Discussion

H1 was supported. H2 received partial support. H3 was partially supported. We found that usage volume alone poorly predicted AI-prohibited exam performance, supporting H1. The CAI metric showed limited overall predictive strength (pseudo-R^2 = 0.138), partially supporting H2. Its weak overall correlations with exam performance (CAI-A: r = − 0.109; CAI-B: r = − 0.047) suggest that aggregating AI interaction behaviors into a single autonomy score may obscure important phase-specific relationships between AI use and academic performance. Moreover, PCA-based models explained more variance than usage volume alone (pseudo-R^2 = 0.140). Critically, phase-specific analyses supported H3: Early conceptual inquiry was strongly associated with higher exam performance. Mid-course contextual questioning showed an initial negative association that was not robust to outlier removal; after excluding the outlier case, no general relationship remained. Late-stage provide-context behaviors were negatively associated with exam performance. Results remained consistent across different statistical approaches (OLS vs. beta regression) and CAI definitions (conservative CAI-A vs. stricter CAI-B). Notably, CAI-B correctly classifies 'provide context' as negative, while CAI-A does not.

The overall model fit, which is a pseudo-R^2 value of 0.140, is not very high. This level is typical in behavioral research since exam results are influenced by numerous unquantified variables, including desire, prior knowledge, and the quality of instruction. Even so, the phase-specific effects are practically meaningful. A 0.10 rise in early conceptual questioning forecasts a 10.8% increase in exam success likelihood. A 0.10 increase in late “provide context” predicts a 15.9% lower chance of passing the exam. These effects were not uniform across the learning process. In the early and late phases, the relationships were stable and aligned with the expected pattern. This suggests that the impact of AI

interaction patterns depends on the stage of learning. These results suggest that the frequency of AI use has a limited relationship with educational outcomes. There was only a weak association between total turns and test scores ($r = .130$, $p = .175$). These findings are consistent with recent research in where the educational impact of generative AI is determined by the quality of learner engagement [31]. On the other hand, behavior impacts that were specific to each phase were important. This outcome is consistent with depth-of-processing theory [32]. CAI translates ideas about self-regulated learning into patterns of conversation that can be measured.

Educators should employ tactics that are specific to each phase.

- **Early Course (Weeks 1–5)**: Use organized prompts to get students to think about concepts. For example, ask ChatGPT to explain the 'why' behind a concept. This pattern was associated with stronger foundational exam performance. Writing requests were associated to greater performance on early tests. This trend shows that giving students little aid at the beginning can help them get started. In later stages, however, heavy reliance on AI-generated solutions may reduce opportunities for independent problem-solving practice.
- **Mid-Course (Weeks 6–10)**: Generally, mid-course contextual questions were found not to have any reliable association with exam performance. The effect of such questions during skill consolidation is not clear, possibly because there are other factors (e.g., prior knowledge, motivation) that are more important and not considered in our models.
- **Late Course (Weeks 11–15)**: Educators should not let students dump text from their assignments. Design assessments that require synthesis rather than reproduction. This reduces the incentive for passive outsourcing.

AI tutoring systems could adapt prompts based on course phase (concept-focused early, synthesis-focused late).

7. Conclusion and future work

This study examines a quality-driven framework for AI integration in education. By moving beyond how often students use AI to assess how they engage with it, we offer educators practical insights. When students use AI matters as much as how they use it. Asking AI for code examples helped students early in the course but would likely hurt later on, when independent problem-solving matters more. The CAI provides a useful way to talk about these interactions, but trying to capture everything in one score hides these important differences. This pattern suggests that what is associated with early learning may, in some cases, be negatively associated with later learning. Educators should therefore focus on whether students' AI use fits where they are in the course.

In this study, some limitations warrant consideration. Unobserved confounders (prior knowledge, motivation) may influence both AI behaviors and exam performance. While our models explain 14% of the variance in outcomes, the majority of variance remains unexplained, and causal claims would be

premature. Another limitation is that this sample is from a course on AI. In this situation, students may be more skilled or more motivated to use AI than students in other areas. It needs to be able to replicate this work in fields like the humanities and the natural sciences. Moreover, the hierarchical dialogue-act labels have been checked, but there is still a chance of labeling noise. This risk is greater for unclear categories like “provide context”. Another limitation is that AI-prohibited exams may not fully capture the types of learning fostered by AI collaboration. Future studies should include both independent and collaborative assessments. Sample selection bias may also have affected our results. The analytic sample includes 110 students, about 54% of the original StudyChat cohort. Participation required consent to share interaction and assessment data, which may introduce self-selection bias if consenting students differ in engagement, performance, or attitudes toward AI.

Future studies could address this limitation using experimental designs that randomly assign students to different AI-use conditions. Future research could also monitor CAI scores across different courses, educational systems, and cultural settings to determine whether our findings generalize to other contexts. Future work could analyze how students write their prompts, such as sentence complexity or emotional tone, to better understand their thinking.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GenAI / GAI	Generative AI
LLM	Large Language Model
CAI	Cognitive Autonomy Index
CAI-A	CAI, Conservative variant
CAI-B	CAI, Stricter variant
POS	Autonomy-supportive (positive) behaviors
NEG	Autonomy-undermining (negative) behaviors
ICAP	Interactive, Constructive, Active, Passive (framework)
CHAT-ACTS	Pedagogical chatbot framework (cited framework name)
PCA	Principal Component Analysis
PC1 / PC2 / PC3	Principal Component 1 / 2 / 3
OLS	Ordinary Least Squares
AIC	Akaike Information Criterion
R^2 / pseudo-R^2	Coefficient of determination / pseudo coefficient of determination
VIF	Variance Inflation Factor
SRL	Self-Regulated Learning
SD	Standard Deviation
SE	Standard Error
OR	Odds Ratio
RQ1–RQ4	Research Question 1–4
H1 / H2 / H3	Hypothesis 1 / 2 / 3
e1, e2, e3	Exam 1, 2, 3 (AI-prohibited)
a1–a7	Assignments 1–7 (AI-allowed)
no_llm / llm	Mean of AI-prohibited exams / mean of AI-allowed assignments
CSV	Comma-Separated Values
JSONL (.jsonl)	JSON Lines
UK	United Kingdom
ε (epsilon)	Boundary-adjustment constant (0.001)

β	Beta coefficient (log-odds)

Declarations

Funding:
This research received no external funding.

Institutional Review Board Statement:
This study analyzed a publicly available, de-identified dataset (StudyChat, McNichols et al. [8]) licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Therefore, institutional review board approval was not required. All data were previously anonymized by the original data collectors.

Informed Consent Statement:
Not applicable.

Data Availability Statement:
This study used the StudyChat Dataset created by McNichols et al. [8], which is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The dataset is available at: https://huggingface.co/datasets/wmcnicho/StudyChat

Acknowledgments:
During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for language editing, sentence rephrasing, and improving readability. The authors reviewed and edited all outputs and take full responsibility for the content of this publication.

Conflicts of Interest:
The authors declare no conflicts of interest.

Clinical trial number:

Not applicable.

Ethical approval:

Not applicable.

Consent to participate:

Not applicable.

Consent to publish:

Not applicable.

References

1. Freeman J. *Student Generative AI Survey 2025*. London (UK): Higher Education Policy Institute (HEPI); 2025 Feb 26. Available from: https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/
2. Vieriu AM, Petrea G. The impact of artificial intelligence (AI) on students' academic development. Educ Sci. 2025;15(3):343. https://doi.org/10.3390/educsci15030343
3. Guo F, Zhang L, Shi T, Coates H. Whether and when could generative AI improve college student learning engagement? *Behav Sci.* 2025;15(8):1011. https://doi.org/10.3390/bs15081011.
4. Cotton DR, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. *Innov Educ Teach Int.* 2024;61(2):228–239. https://doi.org/10.1080/14703297.2023.2175028.
5. Susnjak T, McIntosh TR. ChatGPT: the end of online exam integrity? *Educ Sci.* 2024;14(7):656. https://doi.org/10.3390/educsci14070656.
6. Zimmerman BJ. Becoming a self-regulated learner: an overview. *Theory Pract.* 2002;41(2):64–70. https://doi.org/10.1207/s15430421tip4102_2
7. Stefanou CR, Perencevich KC, DiCintio M, Turner JC. Supporting autonomy in the classroom: ways teachers encourage student decision making and ownership. *Educ Psychol.* 2004;39(2):97–110. https://doi.org/10.1207/s15326985ep3902_2
8. McNichols H, Ikram F, Lan A. The StudyChat Dataset: analyzing student dialogues with ChatGPT in an artificial intelligence course. In: *Proceedings of the 16th International Learning Analytics and Knowledge Conference (LAK '26)*. Kyoto, Japan; 2026. p. 53–63.
9. Bauer E, Greiff S, Graesser AC, Scheiter K, Sailer M. Looking beyond the hype: understanding the effects of AI on learning. *Educ Psychol Rev.* 2025;37:45. https://doi.org/10.1007/s10648-025-10020-8
10. Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, et al. Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br J Educ Technol.* 2025;56(2):489–530. https://doi.org/10.1111/bjet.13544
11. Maitra A, French D, von der Wense K. Dialogue acts as a lens on human–LLM interaction: analyzing conversational norms in model-generated responses. In: *Proceedings of the Fourth Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. November 2025; p. 317.
12. Bašić Ž, Banovac A, Kružić I, Jerković I. ChatGPT-3.5 as writing assistance in students' essays. *Humanit Soc Sci Commun.* 2023;10:750. https://doi.org/10.1057/s41599-023-02269-7.
13. UNESCO. *Guidance for generative AI in education and research*. Paris (France): United Nations Educational, Scientific and Cultural Organization (UNESCO); 2023. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000386693
14. Estaphan S, Kramer D, Witchel HJ. Navigating the frontier of AI-assisted student assignments: challenges, skills, and solutions. *Adv Physiol Educ.* 2025;49(3):633–639.

https://doi.org/10.1152/advan.00253.2024.

15. Ifenthaler D, Schumacher C. Reciprocal issues of artificial and human intelligence in education. *J Res Technol Educ.* 2023;55(1):1–6. https://doi.org/10.1080/15391523.2022.2120257.
16. Stojanov A, Liu Q, Koh JHL. University students' self-reported reliance on ChatGPT for learning: a latent profile analysis. *Comput Educ Artif Intell.* 2024;6:100243. https://doi.org/10.1016/j.caeai.2024.100243.
17. Chi MTH, Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ Psychol.* 2014;49(4):219–243. https://doi.org/10.1080/00461520.2014.965823.
18. Pallant JL, Blijlevens J, Campbell A, Jopp R. Mastering knowledge: the impact of generative AI on student learning outcomes. *Stud High Educ.* 2025;50:1–22.
19. Shi J, Liu W, Hu K. Exploring how AI literacy and self-regulated learning relate to student writing performance and well-being in generative AI-supported higher education. *Behav Sci.* 2025;15(8):705. https://doi.org/10.3390/bs15080705.
20. Rahman S. & Alam A. The behavioral determinants of generative AI use and self-directed learning outcomes in higher education. Discov Educ (2026). https://doi.org/10.1007/s44217-026-01681-w
21. **Wu XY, Chiu TKF.** Integrating learner characteristics and generative AI affordances to enhance self-regulated learning: a configurational analysis. *J New Approaches Educ Res*. 2025;14(1):10. https://doi.org/10.1007/s44322-025-00028-x
22. Lim, T. P., & Teh, W. C. (2026). IT mindfulness and AI literacy shape lifelong learning orientation through student engagement and learning agency in AI-enhanced online learning. Discover Education, 5(1). https://doi.org/10.1007/s44217-026-01219-0
23. Lee HP, Sarkar A, Tankelevitch L, Drosos I, Rintel S, Banks R, Wilson N. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan; 26 Apr–1 May 2025; p. 1–22.
24. Zhang M. Optimizing academic engagement and mental health through AI: an experimental study on LLM integration in higher education. *Front Psychol.* 2025;16:1641212. https://doi.org/10.3389/fpsyg.2025.1641212.
25. de la Puente M, Torres J, Troncoso ALB, Meza YYH, Carrascal JXM. Investigating the use of ChatGPT as a tool for enhancing critical thinking and argumentation skills in international relations debates among undergraduate students. *Smart Learn Environ.* 2024;11:55. https://doi.org/10.1186/s40561-024-00303-2.
26. Lin MPC, Chang D. CHAT-ACTS: a pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning. *Comput Educ Artif Intell.* 2023;5:100167. https://doi.org/10.1016/j.caeai.2023.100167.
27. Zhan Y, Yan Z. Students' engagement with ChatGPT feedback: implications for student feedback literacy in the context of generative artificial intelligence. *Assess Eval High Educ.* 2025:1–14. https://doi.org/10.1080/02602938.2025.2471821

28. Xiao C, Xu SX, Zhang K, Wang Y, Xia L. Evaluating reading comprehension exercises generated by LLMs: a showcase of ChatGPT in education applications. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Toronto (Canada); July 2023; p. 610–625. doi:10.18653/v1/2023.bea-1.52.
29. Sparrow B, Liu J, Wegner DM. Google effects on memory: cognitive consequences of having information at our fingertips. *Science.* 2011;333(6043):776–778. https://doi.org/10.1126/science.1207745.
30. Scarlatos A, Lee J, Woodhead S, Lan A. Simulated students in tutoring dialogues: substance or illusion? *arXiv* [Preprint]. 2026. arXiv:2601.04025. Available from: https://arxiv.org/abs/2601.04025
31. Zhou M, Peng S. The usage of AI in teaching and students' creativity: the mediating role of learning engagement and the moderating role of AI literacy. Behav Sci. 2025;15(5):587. https://doi.org/10.3390/bs15050587
32. Craik FIM, Lockhart RS. Levels of processing: a framework for memory research. *J Verbal Learn Verbal Behav.* 1972;11(6):671–684. https://doi.org/10.1016/S0022-5371(72)80001-X.

Figures

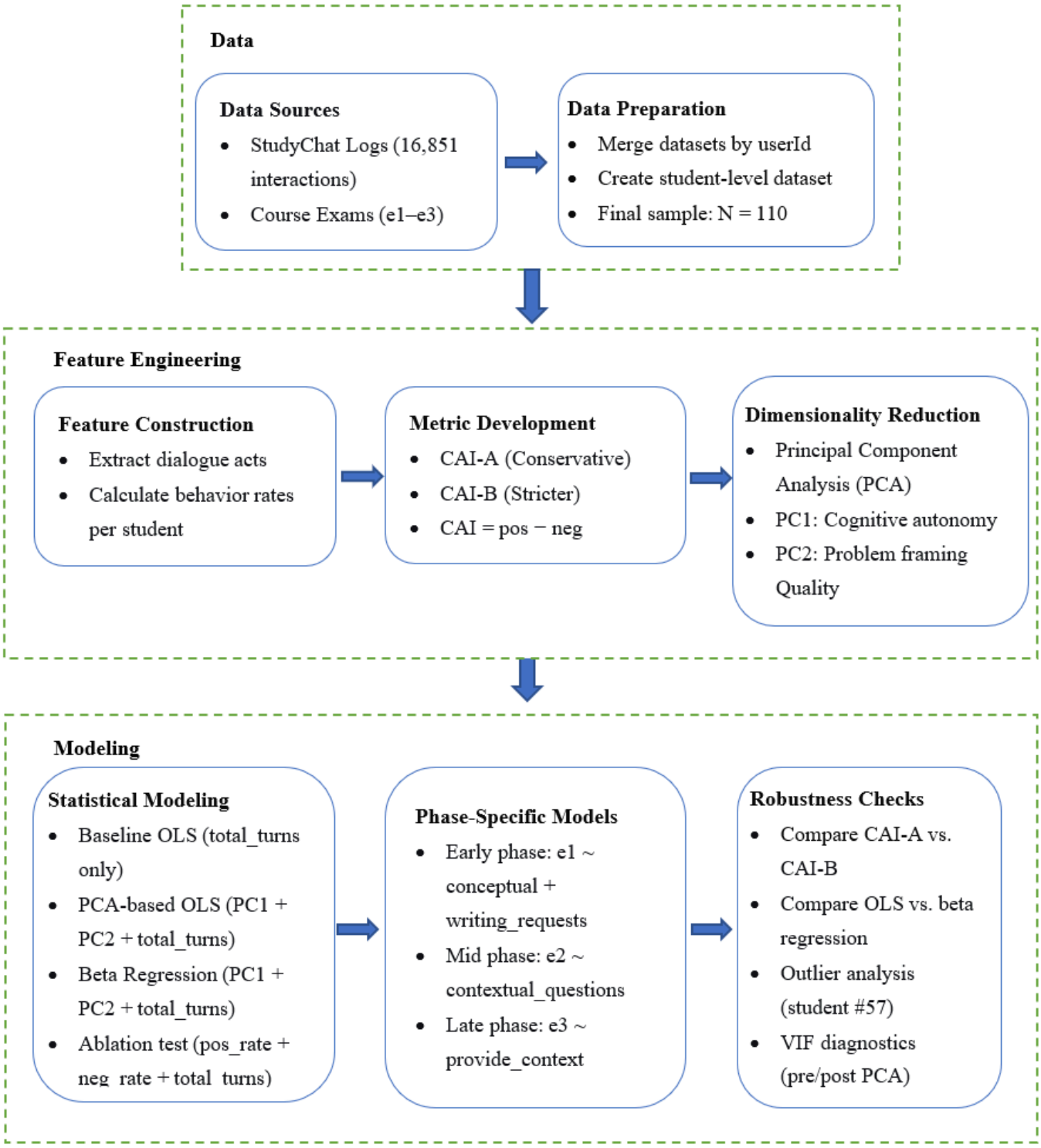


Figure 1

Methodology Steps

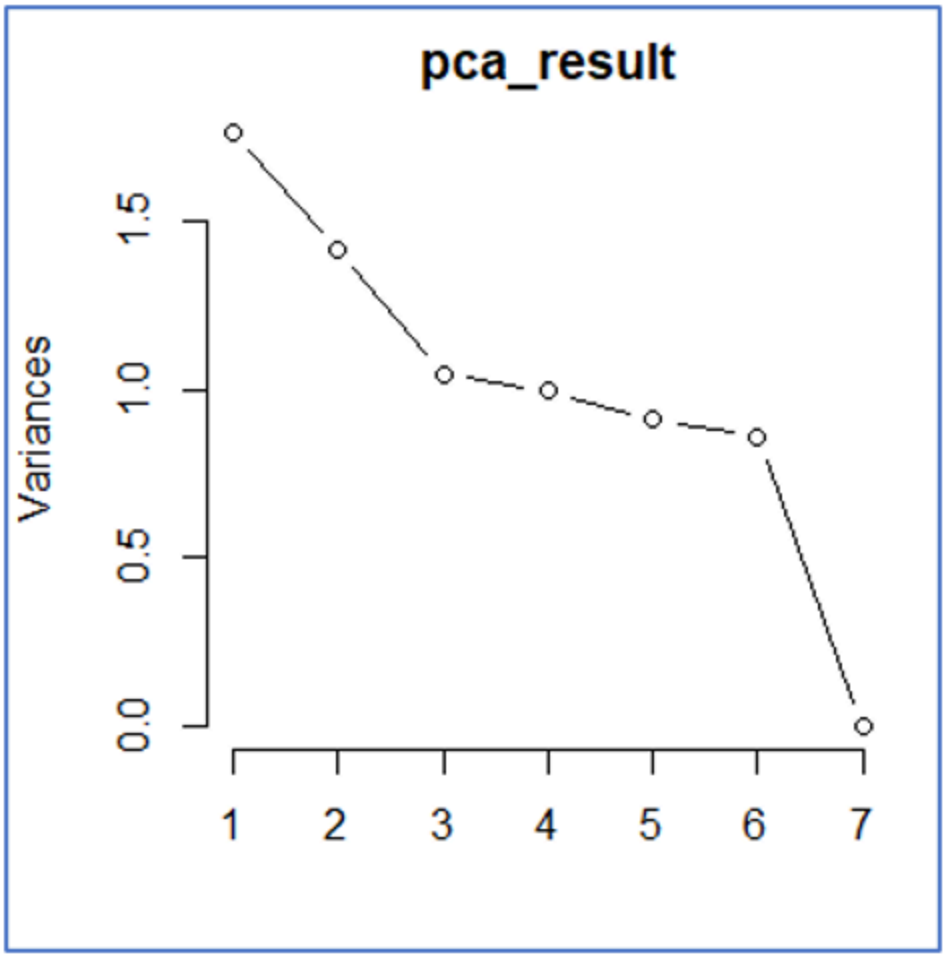


Figure 2

Scree Plot

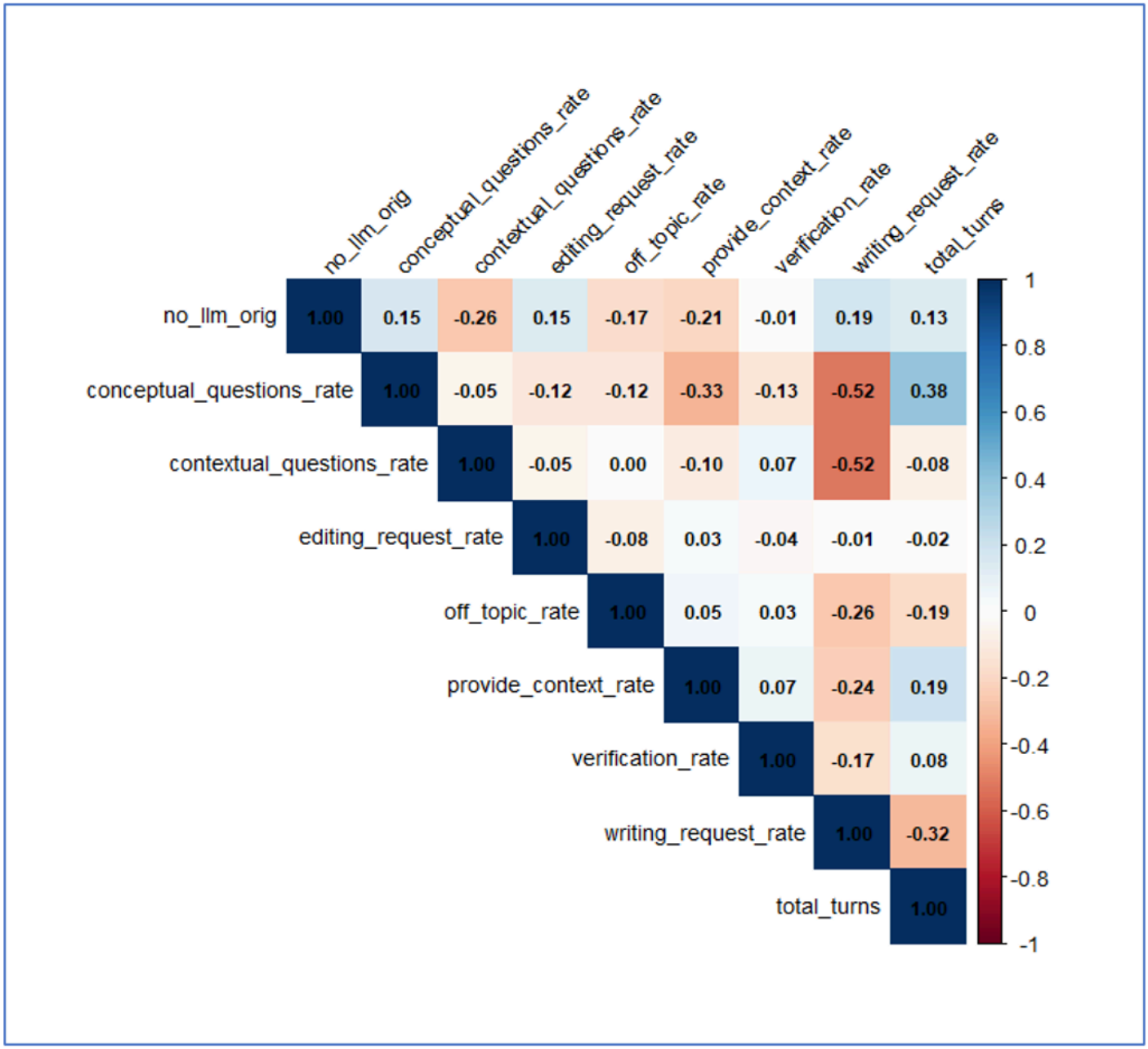


Figure 3

Correlation Matrix Heatmap