Using LLMs to Improve Student Outcomes

Advisor: Dr. Sesh Murthy

Timothy Harmon & Clinton Anderson

Abstract

Large Language Models (LLMs), such as ChatGPT, show promising results for enhancing overall education, but the evaluation of their impact on student performance requires realistic data and rigorous analysis. This project, “Using LLMs to Improve Student Outcomes,” addresses these challenges in two phases.

For the first phase, **Cohort Data Analysis**, rigorous assessment was made to decide whether using an AI helper (LLM-based tutor) in the Math 3B course improves student performance compared to traditional instruction. Using student information (e.g., Test scores, lab assignments, message counts to the AI, etc.) provided by Vocareum, their online learning platform, the students who received AI tutoring were academically matched with similar peers who had no instruction with the AI, allowing one to draw causal insights on the AI tool’s impact.

For the second phase, **Vocareum AI Helper Developmen**t, the goal was to improve and evaluate the online learning platform’s AI helper system. This phase involves analyzing the AI helper’s performance using metrics such as message counts, usage type, and frequency of usage. Successful demonstration produced charts, graphs, and other metrics, with a focus on building a testable, reproducible pipeline to simulate student learning scenarios, enabling analysis of the AI’s feedback quality and identifying any needed enhancements.

Introduction and Question Formulation

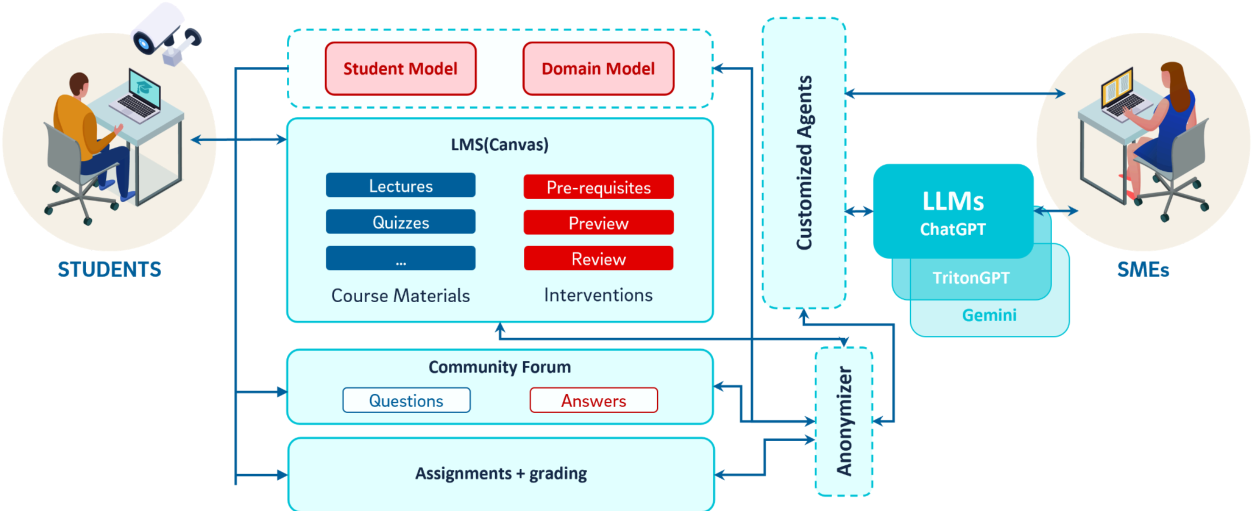


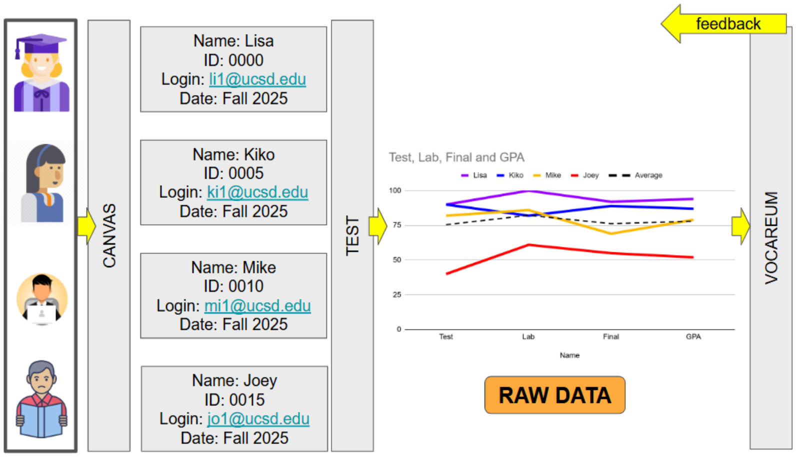
Figure 1. Original pipeline (“Hypatia”) from Dr. Sesh Murthy (w/addition of the Anonymizer)

Challenge (at Start)

Large Language Models (LLMs) can provide personalized assessment and feedback based on students' learning performance. By analyzing students' answers, understanding levels, and error patterns during the learning process, LLMs can deliver targeted assessment results and actionable improvement suggestions. However, while LLMs show promise, they are fragile and fallible, often prone to making errors. Detecting and mitigating these mistakes requires specialized expertise, posing a challenge in their implementation for education.

Problem Statement (Present)

Educators lack data-driven insight into how student interactions with an LLM-based AI Tutor translate into improved learning outcomes. The project was modified to address this gap by analyzing real, anonymized student data from UCSD’s Math 3B course and correlating usage patterns with academic performance.



Name: Lisa
ID: 0000
Login: li1@ucsd.edu
Date: Fall 2025

Name: Kiko
ID: 0005
Login: ki1@ucsd.edu
Date: Fall 2025

Name: Mike
ID: 0010
Login: mi1@ucsd.edu
Date: Fall 2025

Name: Joey
ID: 0015
Login: jo1@ucsd.edu
Date: Fall 2025

Figure 2. Original Pipeline Diagram (before the Anonymizer was implemented)

A typical class is shown in Figure 2. Students can range in capability from top-scoring students who have all prerequisites clearly understood to ones who need greater assistance to succeed. Vocareum is an online-learning platform that seeks to bridge this gap by collecting statistical information about each student and their study habits while also sending those students to an LLM to provide on-demand feedback to learners and to promote greater involvement from instructors. The project evolved to answer two questions:

Do students perform significantly better when using Vocareum?

Can development be done to better comprehend the vast amount of data retrieved from Vocareum?

Team Roles and Responsibilities

Timothy Harmon - Administration Security
Mr. Harmon worked on the security of the project to ensure compliance with the Family Educational Rights and Privacy Act (FERPA) and the main visualizations for the project via Tableau. Mr. Harmon, a distinguished cybersecurity professional with master's degrees in Computer Science and Cyber Security, has established himself as a leader in the field through his pivotal roles as Security Analyst and IT Technician. His deep expertise in threat detection and security analysis is complemented by his active participation in Capture the Flag (CTF) competitions and membership in the Cisco Insider Champion Program, demonstrating his ongoing commitment to practical skill development and industry engagement. Mr. Harmon's blend of technical acumen, educational mentorship, and community involvement makes him an exceptional asset to the cybersecurity landscape, continuously contributing to the field's advancement through his professional work and extracurricular activities.

Clinton Anderson - Design and Development
Mr. Anderson worked on parsing the data, data exploration, and visualizations. He currently works as a Scientist at the Naval Information Warfare Center Pacific, with a research focus on human-computer interaction and learning design. Mr. Anderson holds degrees in Computer Science with an Education minor, a Master's in Game Development, and a Master's in Cybersecurity. His background in education includes teaching English in Japan and introductory computer science at Pasadena City College, reinforcing his interest in optimizing learning systems through technology.

Data Acquisition

Data Sources

Grades-3B-May12-SP2.csv has the Tests from HW 1-4 as well as the Midterm 1 grades, among many other unused headers that were removed. Math 3B Anonymized Gradebook.csv provides 8 Homeworks Labs. The most important correlating factor is that they both have ‘Student #’, which is critical to merge the two tables. It also has “Random ID” which shows a generated number for people who logged on to the AI at least once, and was marked “#N/A” for people who did not. This allows for splitting the table into those two types of users. These anonymous headers were relabeled as “Vocareum ID” across all files for shortness and clarity.

The raw output of the CSVs was imported and cleaned in a Google Colab Notebook. This allows the data to change–for instance, as the semester progresses with additional homework assignments, another Midterm, or even the Final, the pipeline remains adaptable for updated results with minimal modification. Some columns were hardcoded (e.g., columns=[‘HW 1’, ‘HW 2’, …]), but a developer who is versed in data science can add or change these fields easily, maintaining scalability.

To make the data more robust, headings that were seemingly written ‘on-the-fly’ were addressed. Ideally, a Vocareum Administrator should be able to press a button and have a standard and expected CSV result in a given format. However, names like “From Ko”, meaning from the Administrator named Ko, suggests that this was manually populated. Despite this, markdown instructions in the notebook are clear, allowing for common and easy adjustments. Therefore, this pipeline can be recreated for any Vocareum class during any timeframe of the semester with minimal effort.

Grades-3B-May12-SP2.csv and Math 3B Anonymized Gradebook.csv were provided from Vocareum with proper anonymization techniques (no names or PII), downloading the items as an Excel sheet, then manually converting them to CSVs. The other 6 CSVs beginning with “Capstone” were also retrieved from a Vocareum administrator and stored on a Google Drive location with provided access (“Anonymized Math 3B Tutor Conversations Week 1-4.xlsx”). All files were stored online via Google Drive–no database or external cloud storage was used.

Python Notebooks used included Data Cleaning to create usable data frames for user splitting and exploration, and Data Comparisons for initial graphs based on the Labs and Tests. The cleaned data went further on to Tableau. When combined with 6 other CSVs, it demonstrates the quality and quantity of students who used Vocareum.

Data Summary

Descriptions of each dataset are as follows:

- Anonymized Math 3B Tutor Conversations Week 1-4.xlsx - Anonymized excel containing multiple tabs, providing various student interactions with Vocareum
- Grades-3B-May12-SP2.csv - Anonymized and graded set of all students and assignments, including max score possible
- Math 3B Anonymized Gradebook.csv - Lab scores up through HW5
- **capstone_key_categories.csv** - Trimming of message categories down to 6 ‘major’ keys
- **capstone_key_counts.csv** - Shows number of questions asked per ‘minor’ key
- **capstone_rawdata_conversations.csv** - Full messages per student with timestamps and chatIDs
- **capstone_rawdata_questionanswer.csv** - Multiple choice questions and their user responses with timestamps
- **capstone_conversations_studentsideon.csv** - Manual input of the major key with timestamps and the beginning string of the student message
- **capstone_conversation_length.csv** - ChatID and the number of student responses. The total messages is x2 due to the AI’s response
- **ai_chatTotal.csv** - Assignments and scores by students visiting the site at least once, providing an ID and total message count
- **no_ai_chatTotal.csv** - Assignments and scores by students who have never visited the site, and thus do not have an ID nor a message count

Dataset Tables

Dataset Name	Anonymized Math 3B Tutor Conversations Week 1-4.xlsx
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	6 CSVs beginning with “Capstone”
Related Code	None
Data Size	2.1 MB

Table 1: Raw Dataset Table for Grades Anonymized Tutor Conversations

Dataset Name	Grades-3B-May12-SP2.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Math 3B Cleaning.ipynb
Related Code	None
Data Size	25 KB

Table 2: Raw Dataset Table for Grades 3B May 12, Spring 2025

Dataset Name	Math 3B Anonymized Gradebook.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Math 3B Cleaning.ipynb
Related Code	None
Data Size	7 KB

Table 3: Raw Dataset Table for Anonymized student labs

Dataset Name	capstone_key_categories.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	855 B

Table 4: Raw Dataset Table for Key Categories

Dataset Name	capstone_key_counts.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	514 B

Table 5: Raw Dataset Table for Key Counts

Dataset Name	capstone_rawdata_conversations.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	2 MB

Table 6: Raw Dataset Table for Conversations

Dataset Name	capstone_rawdata_questionanswer.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	2.8 MB

Table 7: Raw Dataset Table for Questions and Answers

Dataset Name	capstone_conversations_studentsideon.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	849 KB

Table 8: Raw Dataset Table for Chat IDs, timestamps, message recaps, and keys

Dataset Name	capstone_conversation_length.csv
Input Datasets	Vocareum Teacher Dashboard / Manual Input (Excel)
Destination	Tableau
Related Code	None
Data Size	7 KB

Table 9: Raw Dataset Table for Chat conversations between AI and Student

Processed Data Sets

Dataset Name	**ai_chatTotal.csv**
Source	Grades-3B-May12-SP2.csv & Math 3B Anonymized Student Data.csv
Destination	Math 3B Comparisons.ipynb
Acquisition Code	Math 3B Cleaning.ipynb
Data Size	7 KB

Table 10: Processed Dataset Table for AI with Chat Totals

Dataset Name	**no_ai_chatTotal.csv**
Source	Grades-3B-May12-SP2.csv & Math 3B Anonymized Student Data.csv
Destination	Math 3B Comparisons.ipynb
Acquisition Code	Math 3B Cleaning.ipynb
Data Size	4 KB

Table 11: Processed Dataset Table for No AI with Chat Totals

Dataset Pipeline

Google Colab used here produces the output of the Math 3B Data Cleaning notebook, which splits the input into AI and non-AI dataframes. They are then “(**J**)oined” with the CSVs from Math 3B to make ai_chatTotal and no_ai_chatTotal dataframes. Minimal graphs on the basic input uses Plotly and Seaborn to make basic graphs within Math 3B Data Comparisons. Both data frames are combined with the additional 6 “Capstone” CSVs to make an interactive dashboard in Tableau:

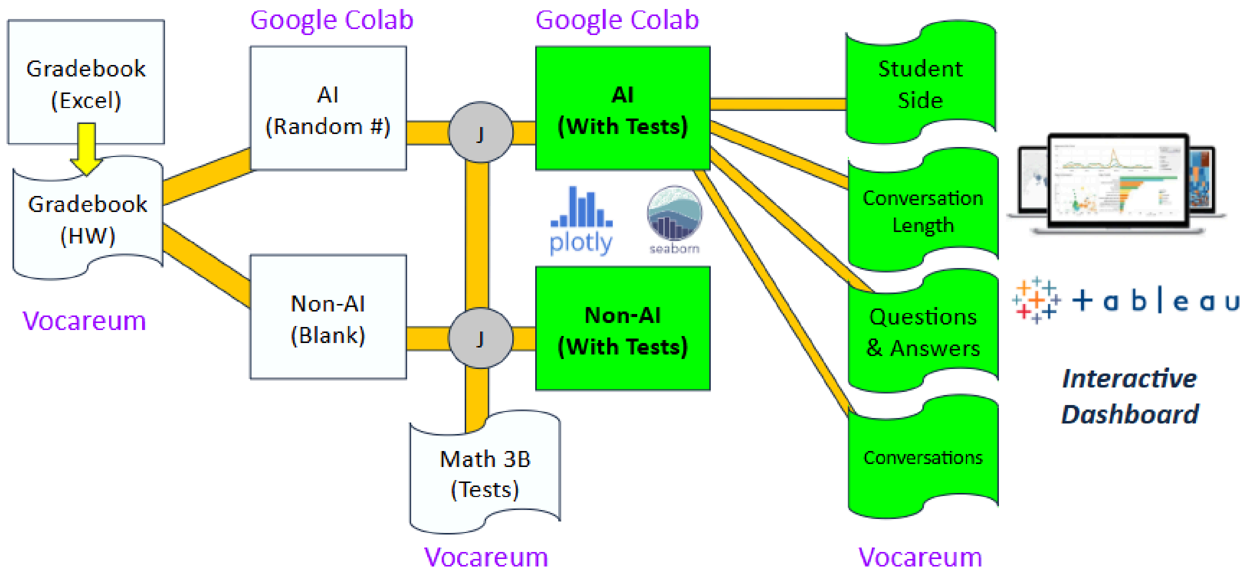


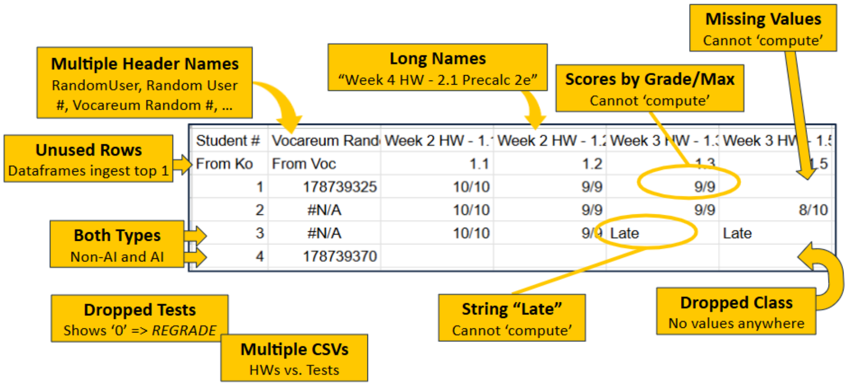
Figure 3. Current Pipeline for Project (Math 3B - other courses follow same pipeline)

Data Preparation

Data Cleaning

The main effort involved in this iteration was to have numerical data and uniform headings, involving:

- Replacing missing data cells with the means of the remaining numerical data
- All scores configured to be numeric (No nulls or strings)
- Shortened names (e.g., ‘Week 4 HW - 2.2 Precalc 2e (963054)’ now “HW 2.2”)
- Ensuring names match and are clear (e.g., the same Vocareum Random #, RandomUser, RandomUser# all transposed to the more readable “Vocareum ID”)
- Ensuring CSV’s are in the standard format (top row header, then data)
- Separation of AI users from non-AI users (with joining *key* of Vocareum ID)
- Transposing “Late” strings to the means of the numerical data
- Dropped Tests (tests scored as “0”) replaced with the mean
- Joining of multiple CSVs *after* splitting into AI users vs. non-AI users
- Removing ‘Dropped students’–those with no data in any column
- Non-inclusion of unused headers (e.g., 1 point updates: “04-02 (959106)”, maximum score fields used in the Notebook calculations but not in the CSV, etc.)



Student #	Vocareum Rand	Week 2 HW - 1.	Week 2 HW - 1.	Week 3 HW - 1.	Week 3 HW - 1.
From Ko	From Voc	1.1	1.2	1.3	1.5
1	178739325	10/10	9/9	9/9	
2	#N/A	10/10	9/9	9/9	8/10
3	#N/A	10/10	9/9	Late	Late
4	178739370				

Multiple Header Names
RandomUser, Random User #, Vocareum Random #, ...

Long Names
“Week 4 HW - 2.1 Precalc 2e”

Scores by Grade/Max
Cannot ‘compute’

Missing Values
Cannot ‘compute’

Unused Rows
Dataframes ingest top 1

Both Types
Non-AI and AI

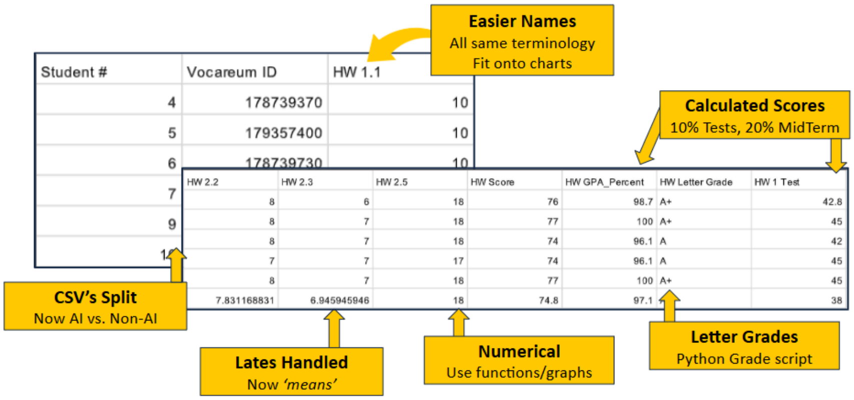
Dropped Tests
Shows ‘0’ => *REGRADE*

Multiple CSVs
HWs vs. Tests

String “Late”
Cannot ‘compute’

Dropped Class
No values anywhere

Figure 4. Example Data before cleaning



Student #	Vocareum ID	HW 1.1
4	178739370	10
5	179357400	10
6	178739730	10
7		
9		
1		

HW 2.2	HW 2.3	HW 2.5	HW Score	HW GPA_Percent	HW Letter Grade	HW 1 Test
8	6	18	76	98.7	A+	42.8
8	7	18	77	100	A+	45
8	7	18	74	96.1	A	42
7	7	17	74	96.1	A	45
8	7	18	77	100	A+	45
7.831168831	6.945945946	18	74.8	97.1		38

Easier Names
All same terminology
Fit onto charts

Calculated Scores
10% Tests, 20% MidTerm

CSV’s Split
Now AI vs. Non-AI

Lates Handled
Now *‘means’*

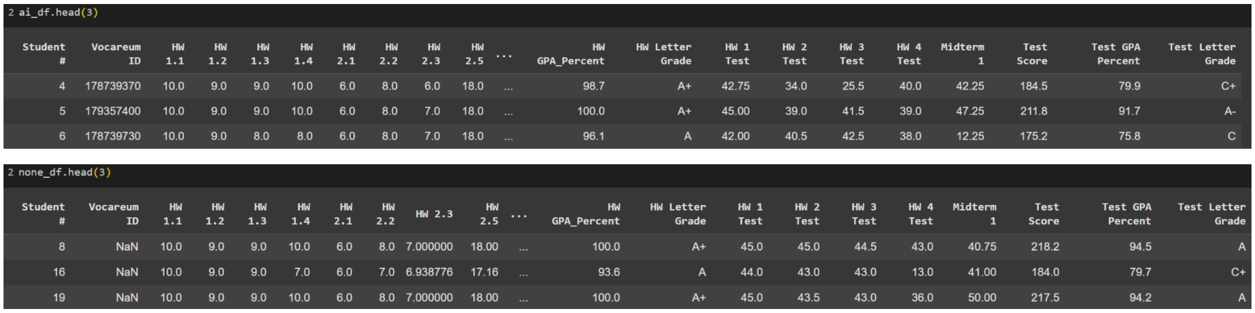
Numerical
Use functions/graphs

Letter Grades
Python Grade script

Figure 5. Example data post cleaning

Data Frames

The results of data cleaning led to two dataframes: **ai_df** and **none_df**:



```
ai_df.head(3)
```

Student #	Vocareum ID	HW 1.1	HW 1.2	HW 1.3	HW 1.4	HW 2.1	HW 2.2	HW 2.3	HW 2.5	...	HW GPA_Percent	HW Letter Grade	HW 1 Test	HW 2 Test	HW 3 Test	HW 4 Test	Midterm 1	Test Score	Test GPA Percent	Test Letter Grade
4	178739370	10.0	9.0	9.0	10.0	6.0	8.0	6.0	18.0	...	98.7	A+	42.75	34.0	25.5	40.0	42.25	184.5	79.9	C+
5	179357400	10.0	9.0	9.0	10.0	6.0	8.0	7.0	18.0	...	100.0	A+	45.00	39.0	41.5	39.0	47.25	211.8	91.7	A-
6	178739730	10.0	9.0	8.0	8.0	6.0	8.0	7.0	18.0	...	96.1	A	42.00	40.5	42.5	38.0	12.25	175.2	75.8	C

```
none_df.head(3)
```

Student #	Vocareum ID	HW 1.1	HW 1.2	HW 1.3	HW 1.4	HW 2.1	HW 2.2	HW 2.3	HW 2.5	...	HW GPA_Percent	HW Letter Grade	HW 1 Test	HW 2 Test	HW 3 Test	HW 4 Test	Midterm 1	Test Score	Test GPA Percent	Test Letter Grade
8	NaN	10.0	9.0	9.0	10.0	6.0	8.0	7.000000	18.00	...	100.0	A+	45.0	45.0	44.5	43.0	40.75	218.2	94.5	A
16	NaN	10.0	9.0	9.0	7.0	6.0	7.0	6.938776	17.16	...	93.6	A	44.0	43.0	43.0	13.0	41.00	184.0	79.7	C+
19	NaN	10.0	9.0	9.0	10.0	6.0	8.0	7.000000	18.00	...	100.0	A+	45.0	43.5	43.0	36.0	50.00	217.5	94.2	A

Figure 6. ai_df and none_df comparing similar metrics

Of 140 students, 103 had complete rows (i.e., all homeworks, tests, and Midterm scores), with 64 as AI users and 39 as non-AI users. The top figure of Figure 6 is 'ai_df', and the bottom one is 'none_df', consisting of these major parts:

- **Student ID** - An anonymized numeral of sequential students provided by Vocareum that shows students as they appear in the provided data.
- **Vocareum ID** - Renamed from "Random #" that shows a student's AI identification. Students without it have blanks and are placed into the *none* df.
- **HW** - From 1.1 to 2.5, consists of 10 point assignments. Both types of students received full credit and have *infinite* attempts, which made for trivial comparison.
- **Tests** - Tests covering HWs 1 - 4 and the Midterm. These have much greater variation and are worth higher points, so they are the main comparison.
- **Scores, GPA Percent, Letter Grade** - Calculated from the addition of either the Homework ("HW…") or Tests ("Test…"). They can be quickly read by a human in this format.

Remaining Data (AI Interaction)

Additional elements from the Anonymized Tutor Conversations capture how students interacted with the AI system throughout the course. These include major question categories (e.g., "Graphing," "Simplification"), usage frequency, conversational behavior with the chatbot, message counts per chat, and total duration per session. This data was excluded from initial Notebook visualizations due to its inapplicability to non-AI users. Moreover, many students posed trivial or redundant queries (e.g., "35-15?", "Is this the right answer?"), which limits the reliability of individual-level analysis. However, aggregate statistics such as total message counts—when compared to final scores—can still yield insight into overall AI usage and effectiveness.

Analysis Methods

Initially, there were 3 selected assignments worth noting, based on the viewed histogram charts and line graphs, between AI and non-AI users: High Midterm 1 scorers (>= 42), low Midterm 1 scorers (< 42), and the HW 4 Test takers.

Comparison using p-tests for these most extreme cases was performed using Scipy's “ttest indicator” library. A t-test demonstrates that statistical means are within a standard deviation of each other, while a p-test indicates the likelihood of observing such a difference by chance, assuming no real difference exists.

An example is shown here, for students scoring high on the Midterm:

```
# Midterm 1 values from both high-scoring groups
t_stat, p_value = ttest_ind(ai_hi['Midterm 1'], none_hi['Midterm 1'],
equal_var=False)
```

The difference is statistically significant when the p-value is less than 0.05 for the two groups:

Hi-scored Midterms	`t-statistic: 1.5576092847583827`
	`p-value: 0.12591439652187655`

Low-scored Midterms	`t-statistic: 0.7829694439105137`
	`p-value: 0.44073482549l8946`

Homework 4 Tests	`t-statistic: 1.0544715308099943`
	`p-value: 0.2967258279546521`

Table 12. Significance observations among the 3 stand-out groups

To contrast, standard variance statistics are as follows (as an example):

- AI-users: `Mean:  32.54 , Std Dev:  8.59 , Error:  1.43`
- Non-AI users: `Mean:  30.21 , Std Dev:  10.79 , Error:  2.62`

The p-values are outside the threshold. Additionally, all groups are within one standard deviation of each other. These values indicate there is no statistical difference between any of the groups.

As such, due to limited exposure to multiple data sources, it made more sense to develop a ‘cleaning and graphical’ pipeline that could be used to verify other data than continue comparing students that performed (basically) the same, for this class in this semester.

Findings and Reporting

A pipeline of cleaned and transparent data with graphical capabilities will be handed off to Vocareum. Not much was surmised from the class itself, with no statistical anomalies detected between the students. However, this case is specific to a single class (Math 3B) for only half of a single semester (Spring 2025). Despite this, the results can be considered a baseline for improvements to the Vocareum platform and to provide guidelines to motivate students to choose to use AI, once improved metrics are adopted based on the findings.

The pipeline aligns well with both academic research and business use cases as well:

- **Educational Research** – Helps UCSD instructors and learning scientists evaluate hybrid instruction models, compare in-person vs. AI-supported outcomes, and identify course pain points.
- **Platform Marketing** – Demonstrates how Vocareum's AI tools drive measurable student engagement and learning improvements. Provides data-backed use cases for institutional adoption.
- **AI Strategy** – Offers metrics to guide future development of Hypatia (UCSD's prior standalone AI tutor) by benchmarking it against Vocareum's usage patterns, message quality, and user outcomes.

Findings from Notebook

The data shows limited improvement in student scores using Vocareum

Key Insights from Final Datasets

- AI group has 64 students, non-AI has 39 (of 140 in set)
- AI group - Higher scores for HW Tests but not Midterm

This negates the prior hypothesis that AI users would perform better.

Techniques Used

- The **describe()** feature shows the Midterm scores are basically the same (~39) for both types of users.
- The fact that the AI users performed consistently better on all the homework tests, sometimes substantially, adds complexity to the interpretation.
- This combined with the **graph** showing lower and fewer high-scores on the Midterm means that AI users do not statistically have an advantage. This was verified via **p-value** from the t-test.

This outcome can be related to a number of factors. Initial considerations are that the AI users were too dependent on the AI for answers, or had less preparation due to not considering hard practice homework questions. Additional data exploration into how the AI system was used using additional information received from CSVs obtained may reveal the root cause.

Notebook Visualizations

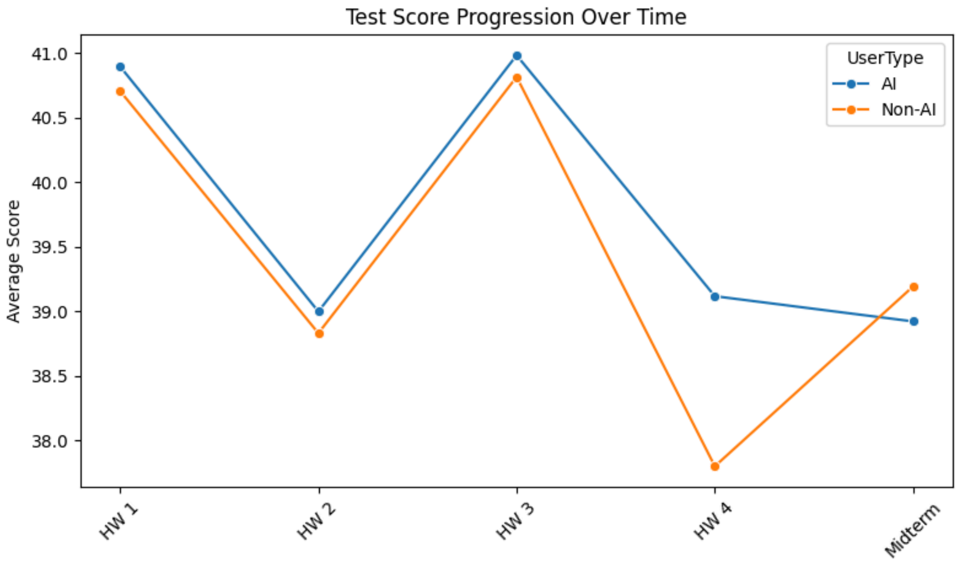


Figure 7. Test Score Progression Over Time (HW 1 through Midterm). It has the interesting result of AI users doing similar (HW 1 Test) or *much* better (HW 4 Test) on tests leading up to the Midterm, but actually performed slightly worse on the Midterm. Tests are performed without AI for all students.

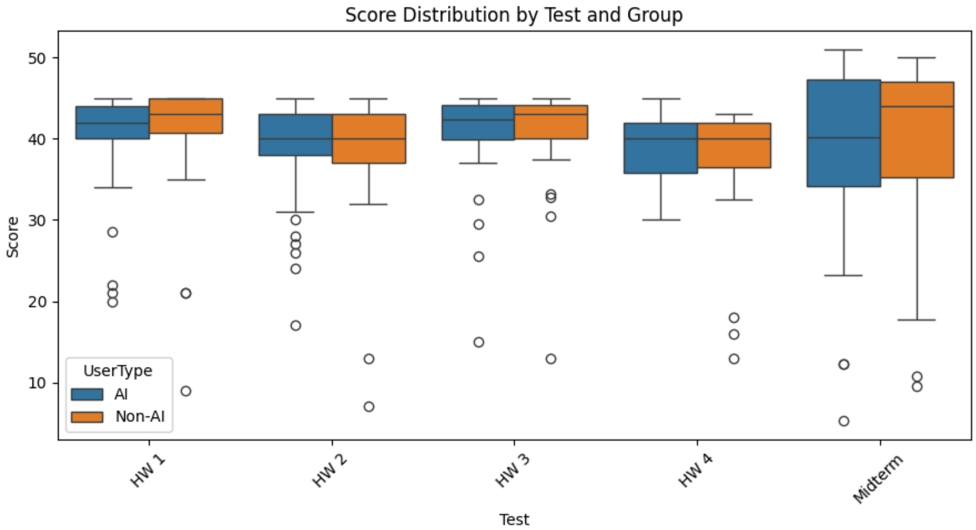


Figure 8. Box Plot showing similar std devs for AI students compared to non-AI. The HW Test 4 non-AI box shows the same median and tighter variation, with the exception of 3 outliers

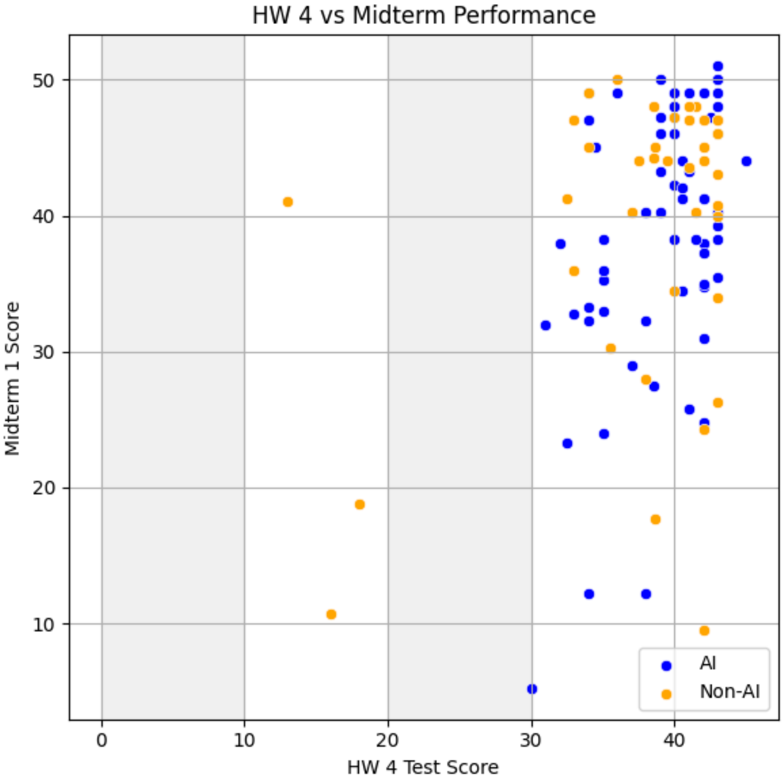


Figure 9. Scatter plot comparing the progression from an earlier Test (HW 4) to the Midterm. 3 low-scoring HW 4 Test non-AI students show different levels of Midterm success. There are mixed types of AI and non-AI students for very low Midterm scorers (< 20).

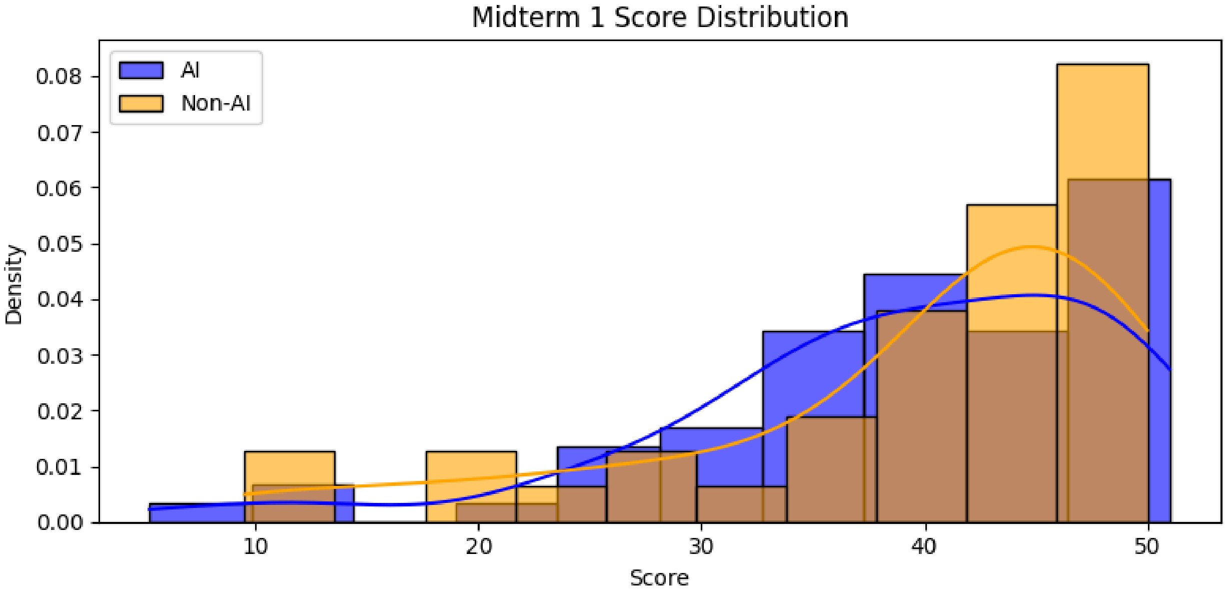
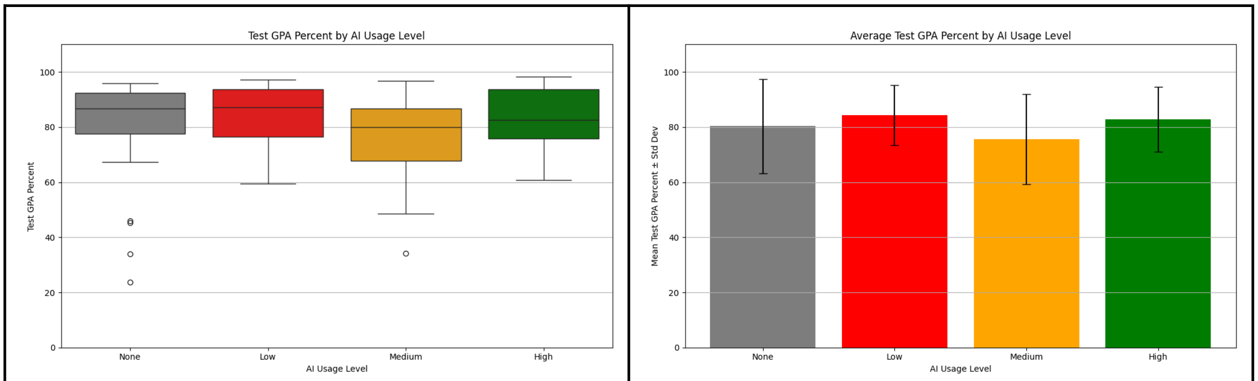


Figure 10. Histogram shows the unexpected result of non-AI students' high comparative scores (~42-46 and 46-51)

A final set of Notebook visualizations centered on test scores, message counts, and student groupings.



a) Box Plot showing **Median** test scores

b) Bar Chart showing Test Score **Means**

Figure 11. Legend: Grey = No AI use (0 messages), Red = Low (1-2 messages), Orange = Medium (3-12 messages), Green = High AI usage (13-100 messages)

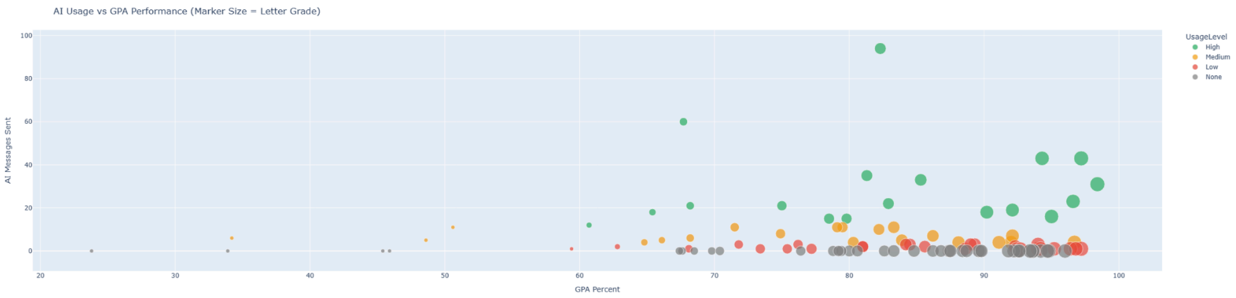


Figure 12. Plotly Scatterplot of GPA scores by AI usage level. At ~50 points and below, instructors could promote AI usage to certain “Grey” (non-user) students, or discover the failure of why the AI was ineffective for the several “Orange” students

Findings from Graphing

Successfully developed end-to-end pipeline culminating in interactive graphical dashboard

Key Insights from Final Dashboard

To demonstrate the success of the new graphical capabilities, several basic dashboards were used to simply explore the capabilities of the cleaned and descriptive data frames. Additional Tableau charts can be examined by Vocareum to provide ad-hoc descriptive analysis of a class:

- **Bar Charts**: Categorical & Numerical comparisons. Letter Grades show similarity for both groups.

- **Sankey** - Categories have the highest count of correct answers, then incorrect, and so on.
- **Heatmap** - Higher AI usage in the evening of Monday & morning of Tuesday than the rest of week.
- **Scatter** - Most students used AI less and received decent grades (A - C-).

Techniques Used

The interactivity allows singling out students, focusing on AI interaction, timeframes, and separating usage statistics into specific sections. The final dashboard can be used with multiple class majors, student sizes, and varied assignments.

Tableau Interactions

Multiple values can be selected by Control-Clicking, or Selection-Square dragging:

- **Heatmap** - Clicking on the day of the week shows how many students studied during that time frame.
- **Message Counter** - Hovering over each plot shows each Letter Grade and the plots for that Letter Grade. Clicking the grade shows the number of students.
- **Total Students** - Clicking AI vs. non-AI bars filters Letter Grades and Message Counter, as well as the Letter Grades by Group, isolating those students.
- **Sankey** - Clicking on a Question Category or an Answer Type shows how many student occurrences happened for that feature.
- **Letter Grades** - Hovering highlights the Letter Grade in Message Counter and the Letter Grade legend. It also highlights the AI Students or non-AI Students bar in the Total Students (AI vs. non-AI) bar graph, and the grade in the filter.

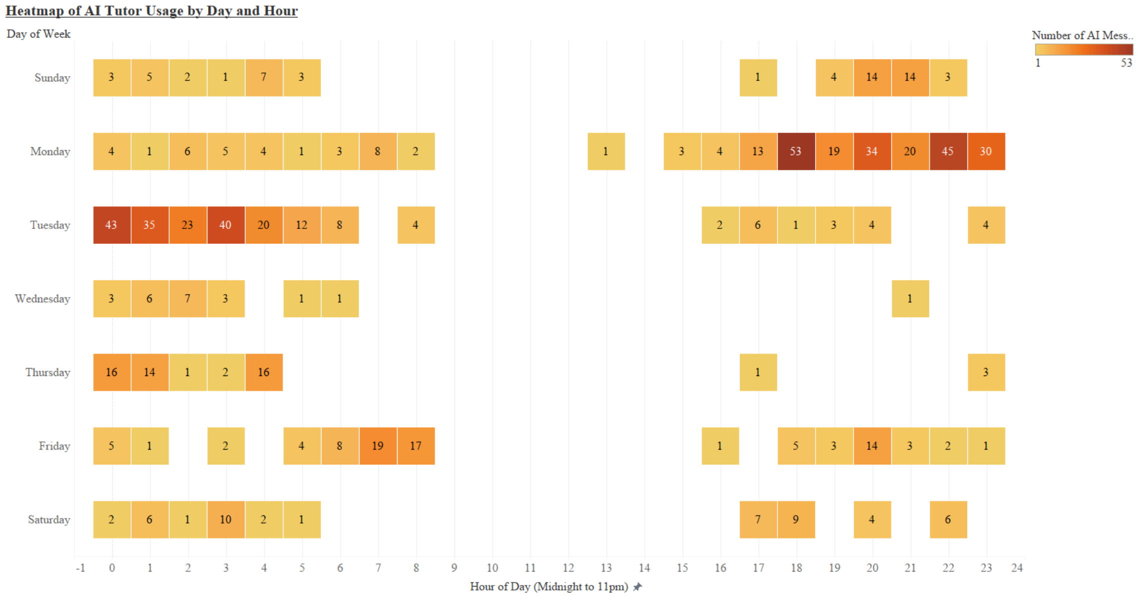
Tableau Filters

Changing of indicators on the right side allows isolated manipulation of multiple charts

- **Student Group** - Highlights Total Students and Letter Grades by Group.
- **Time Granularity** - Filters the heatmap by hour or day.
- **Test Letter Grade** - Highlights students in Message Counter, Letter Grades by Group.

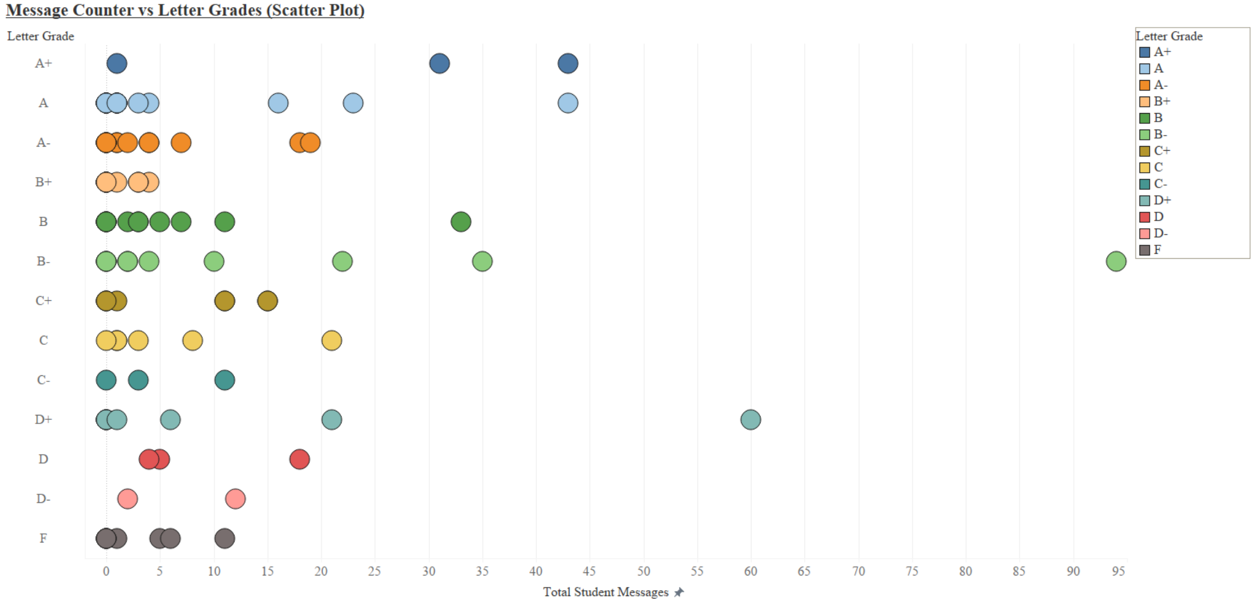
Tableau Visualizations

Visualizations developed in Tableau Desktop show greater variety and customization than the visualizations done via Python libraries. These visualizations have some form of interactivity and range from selecting various parts of the visualization to highlighting the relevant information in order to have the bigger picture.



Hour of Day (Midnight to 11pm) for each Day of Week. Color shows count of capstone_rawdata_conversations.csv. The marks are labeled by count of capstone_rawdata_conversations.csv. The data is filtered on Chat Type and Action (Student Group). The Chat Type filter keeps AI Student and Non-AI Student. The Action (Student Group) filter keeps 2 members.

Figure 13. AI Tutor Usage by Day and Hour (Heatmap) - most AI usage is Monday night into Tuesday morning (this might be because of last minute working on homework due the next day or some other factors).



Student Message Count for each Letter Grade. Color shows details about Letter Grade. Details are shown for Student #. The data is filtered on Student Group, Action (Student Group) and Action (Hour of Day (Midnight to 11pm),Day of Week). The Student Group filter keeps AI Students and Non-AI Students. The Action (Student Group) filter keeps 2 members. The Action (Hour of Day (Midnight to 11pm),Day of Week) filter keeps 108 members. The view is filtered on Letter Grade, which keeps 13 of 13 members.

Figure 14. Message Count vs. Letter Grades (Scatter Plot)

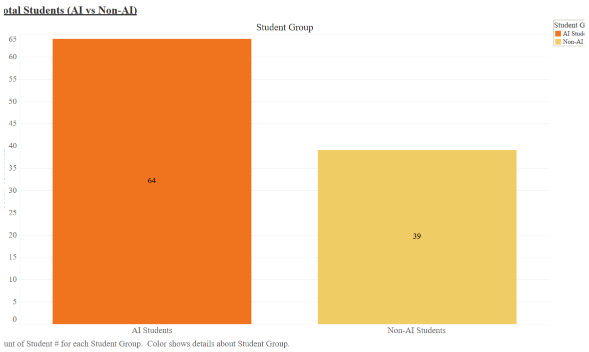


Figure 15. Total Students (AI vs non-AI)

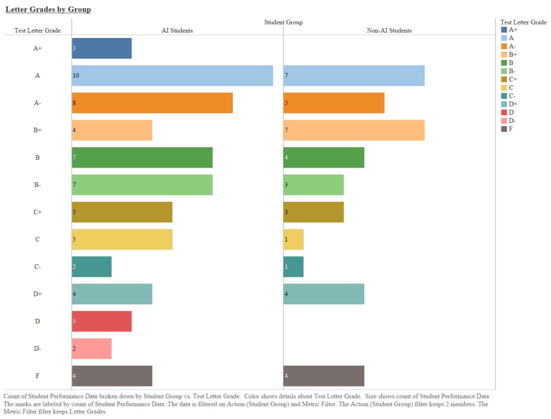


Figure 16. Letter Grades by Group - Bar Chart

Figure 15 shows the number of students who used the AI Tutor vs. who did not use the AI Tutor, whereas Figure 16 shows the breakdown of the Letter Grades for each group (AI Tutor and non-AI Tutor). This interaction is shown in the “AI Tutor Analysis Dashboard,” where clicking on a group dynamically filters the “Letter Grades by Group” chart.

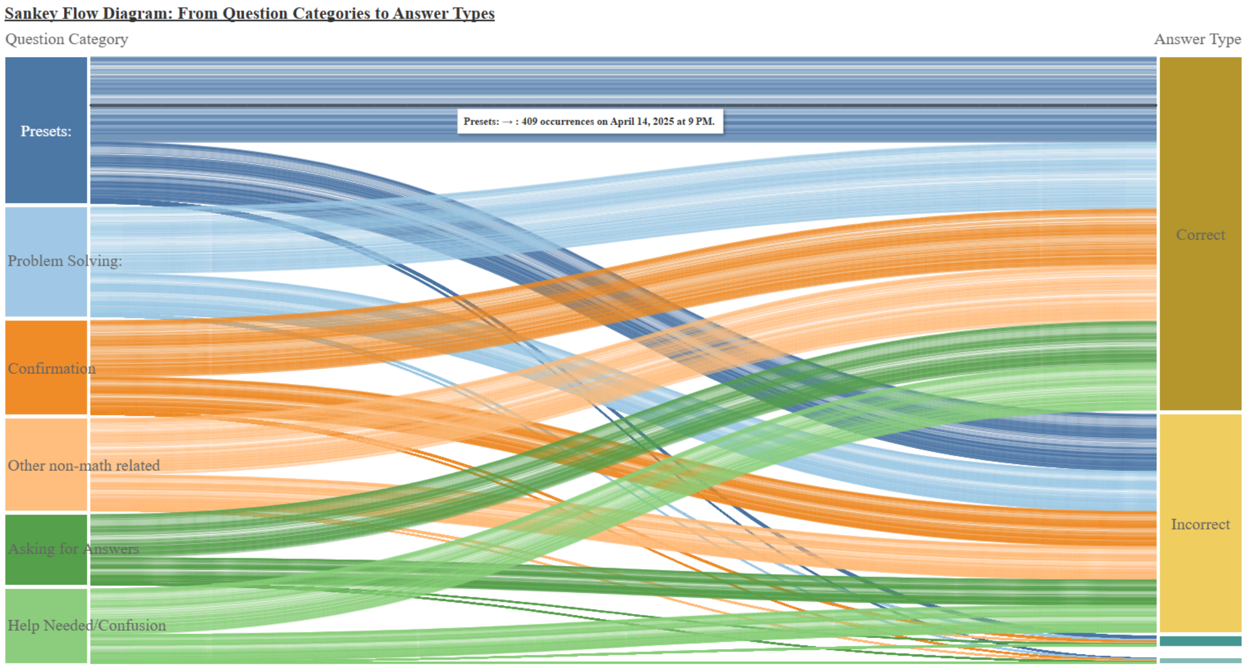


Figure 17. Sankey Flow Diagram: From Question Categories to Answer Types

A Sankey flow diagram is a visualization that is used to depict a flow from one set of values to another. Figure 17 shows a Sankey Flow Diagram for the various categories of questions students answered, how many responses are associated with each category, and each response state (correct, incorrect, skipped, and unanswered).

Tableau Dashboard

Due to Tableau's acquisition by Salesforce, UCSD students lost access to the full Tableau Desktop license. As a workaround, visualizations were packaged as .twbx extracts. These files include embedded data, ensuring dashboards remain functional offline. While this limits public sharing, it enhances data privacy.

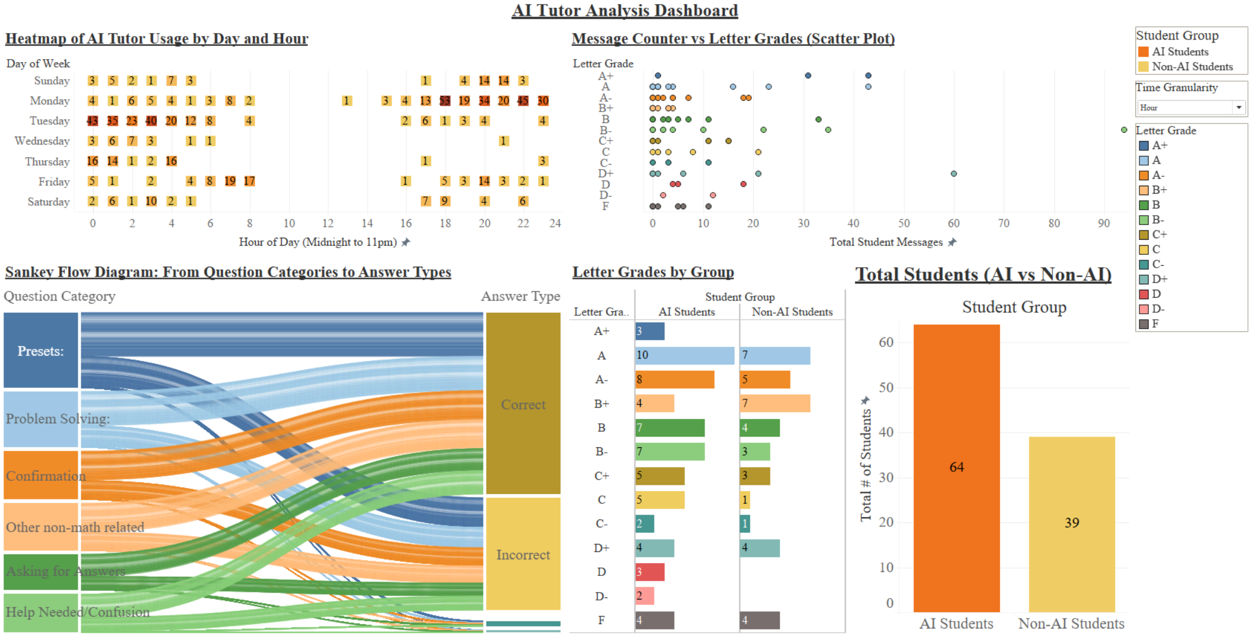


Figure 18. Tableau Dashboard (main part of hypothesis 2)

Solution Architecture, Performance and Evaluation

The final pipeline was validated by applying it to a partially completed course. The system was tested end-to-end—from initial CSV ingestion to interactive visualization—using an earlier session of Math 3B as a proxy for future classes or mid-semester updates.

The evaluation confirmed three key points of utility:

- **Vocareum Administrators** can recreate visualizations for any class or period using the descriptive Tableau dashboards.
- **Student Developers** can modify or extend the pre-built Google Colab notebooks for new courses or test formats.
- **UCSD Researchers** have access to final numerical datasets suitable for deeper statistical analysis, including dispersion measures, normalized comparisons, or cross-dataset correlations.

Capstone stakeholders reviewed final deliverables (notebooks, Tableau dashboards, visuals) and provided positive feedback, confirming that the pipeline met both functional and exploratory expectations.

References

1. UC San Diego Math Placement Exam. *(n.d.)*. *Canvas Course Description*. Retrieved from https://mathtesting.ucsd.edu/placement/mpe

2. Vocareum. *(2025)*. *Instructor Dashboard – Beta Testing Environment for Math 3B*. Internal UCSD Instructor Portal.

3. Harmon, T., & Anderson, C. *(2025)*. *Capstone Drive Notebook*. Shared Google Drive Repository – Raw Data, CSVs, and Jupyter Notebooks.

4. UC San Diego Institutional Review Board. *(2025)*. *ASPIRE – Adaptive Scaffolding for Personalized Instruction and Responsive Education* (IRB #809818). Approved study protocol, UCSD Office of IRB Administration.

5. OpenAI. *(2025)*. *ChatGPT-4o* – Editing, Visualization Support, and Code Assistance. Retrieved from https://chat.openai.com

6. Tableau Software. *(2025)*. *Tableau Public Dashboard – Visual Analytics Platform*. Retrieved from https://public.tableau.com

7. University of California, Santa Barbara. *(2023)*. *Math 3B: Calculus with Applications – Course Syllabus*. Department of Mathematics.

8. Wikipedia Contributors. *(2024, June)*. *Copula (Probability Theory)*. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Copula_(probability_theory)

Appendix A: DSE MAS Knowledge Applied to the Project

The Capstone project, “Using LLMs to Improve Student Outcomes,” integrates core concepts and tools from across the DSE MAS curriculum. Below is a breakdown of how each course contributed to our work:

DSE 200 – Python for Data Analysis
We used Python as our primary programming language, applying pandas, NumPy, and regular expressions for data wrangling and exploratory analysis. Our Jupyter Notebooks formed the foundation of the pipeline, with logic built for cleaning, scoring, and grouping student submissions.

DSE 201 – Database Management Systems
Although we worked mostly with CSVs, our understanding of relational data structures and schema normalization helped us merge multiple datasets (e.g., chat logs, test scores, and student metadata). Our joins and groupings replicated standard SQL practices within pandas.

DSE 203 – Data Integration and ETL
The entire project pipeline reflects an ETL process: raw CSVs from Vocareum were cleaned, normalized, and transformed into analysis-ready formats. Our treatment of message logs, test results, and performance data mimicked the challenges of real-world integration across varied formats.

DSE 210 – Statistics and Probability Using Python
We applied statistical tests such as t-tests to compare performance across groups (AI vs. non-AI, and by scoring levels). We calculated means, standard deviations, and standard errors, and interpreted p-values to assess significance of score differences.

DSE 220 – Machine Learning
While we didn’t train new models, our project simulated the behavior of an LLM tutor. Our analysis focused on evaluating the effectiveness of machine learning-based tools (e.g., the AI tutor in Vocareum) and validating its use via performance trends.

DSE 230 – Scalable Data Analysis
Although our dataset was moderate in size, our preparation of modular and reproducible notebooks ensures scalability for future classes. The use of abstraction and batch analysis practices lays a foundation for expansion.

DSE 241 – Data Visualization
We used Seaborn, Plotly, and Tableau to communicate insights. Our visualization work

incorporated principles of interactivity, clarity, and effective encoding — aligning directly with DSE 241’s focus on storytelling and exploration through charts, dashboards, and interactive filters.

DSE 250 – Beyond Relational Data Models
While not a direct focus, the course’s exposure to graph structures (e.g., “Resource Description Frameworks (RDF) and network-based modeling) inspired us to explore Pyvis (unused, but attempted) and Sankey diagrams to visualize flows between question types and outcomes.

DSE 260 – Capstone Design Project
The entire scope of this project follows the Capstone framework: identifying a domain of interest (AI in education), conducting exploratory and statistical analysis, building a data pipeline, and producing a reproducible final product backed by formal documentation and interactive tools.

DSE 290 – Case Studies in Data Science
Lessons from real-world examples shaped our problem scoping. We focused on practical impact, working closely with stakeholders (faculty, industry, platform developers) to ensure that our deliverables met both academic and operational needs.

Appendix B: Link to UCSD Library Archive

Harmon, Timothy D.; Anderson, Clinton (2025). Using LLMs to Improve Student Outcomes. In Data Science & Engineering Master of Advanced Study (DSE MAS) Capstone Projects. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0RN387D