“Are These ‘Model’ Results?”
Incorporating Large Language Models into Statistical Writing and Coding

Joshua Sparks – Department of Statistics & Data Science, University of California Los Angeles
josh.sparks@stat.ucla.edu

Introduction

- With the onset of OpenAI’s GPT and other Large Language Model (LLM) software in public access, universities continue to navigate the roadmap for student use (and misuse).
- Furthermore, survey results show that students also desire to learn how to ethically incorporate this new technology into their education, as many learners have yet to receive proper training within this landscape.
- Implemented while at a medium-sized private research university, training and assessment was integrated into its writing-intensive, second-year undergraduate course in statistical computing.

Course Details and Student Composition

Description of Course with LLM Topics Implemented

- Sophomore-level course (STA 2183W) in statistical computing, covering: descriptive measures, quantitative and qualitative inference for one or more populations (parametric and nonparametric), data management and visualization, regression and model selection, introductory sampling and simulation
- Topics can be found in an introductory statistics textbook, and there is a prerequisite of an introductory statistics course.
- This course is one of two possible writing-intensive course options within the Department of Statistics.

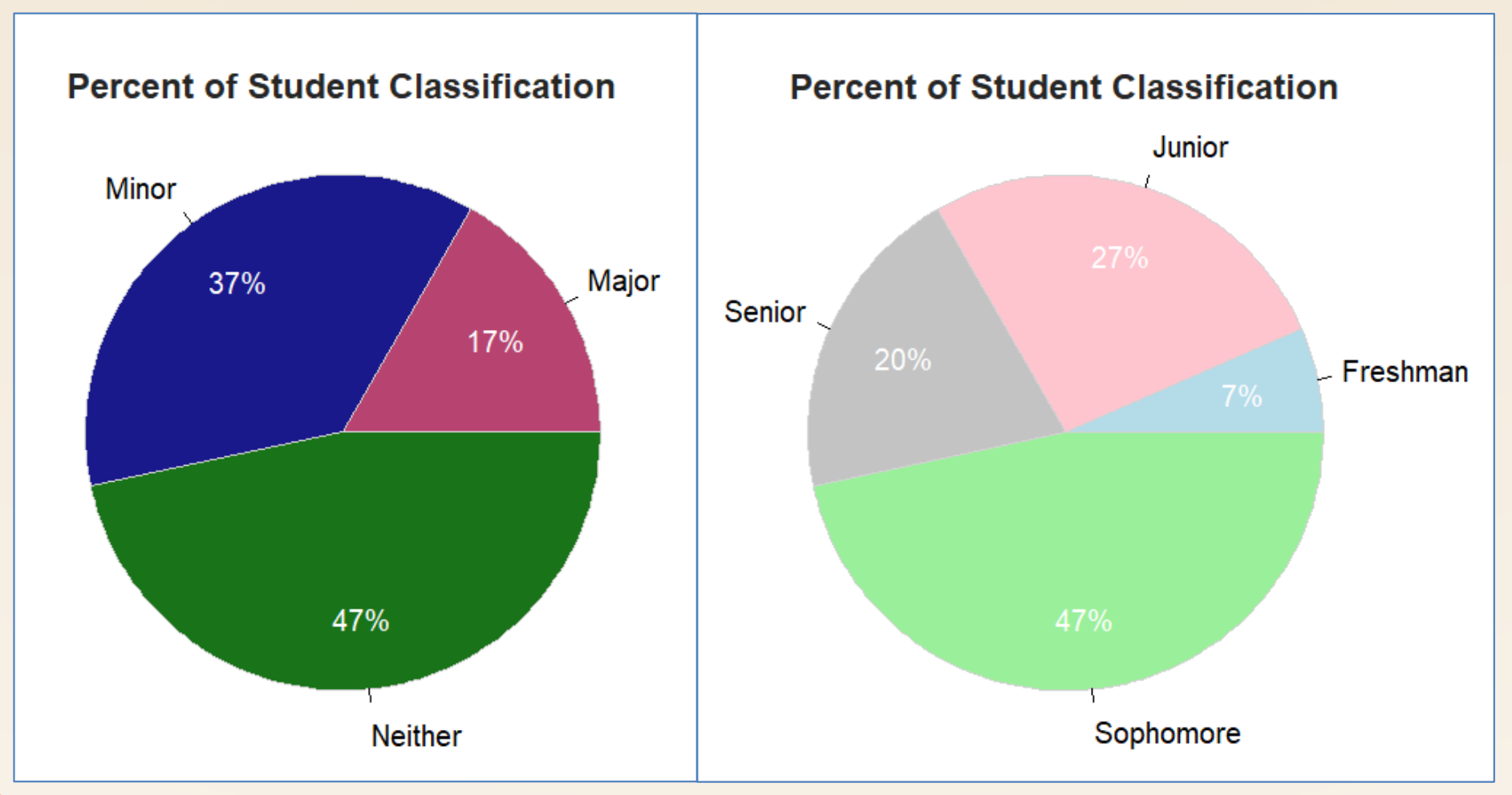
Course Assessment

- Course requirements include:
  (1) weekly homework problems that synthesize the lectures and focus on coding (with comments) and reporting results.
  (2) three consulting reports that include an initial draft, conferencing (Reports 1 and 2), and a final submission.
  (3) two exams that consist of an in-class component resembling homework and a take-home component with a paper-writing portion.

Student Composition

- Roughly 63% of students are taking their first WID course
- Many economics majors (23% of those enrolled) along with engineering, finance, public health, and psychology
- Students vary in statistical preparation, ranging from 1-4 classes before enrolling.
- All students responded that they had no formal training or discussion on AI before (Spring 2024).

Figure 1: Breakdown of Statistics Majors/Minors and Student Classification (n = 30).



Varied Topics Incorporated into Lessons

- Wang et al (2024) discusses the vast literature on AI in education, ranging from tutoring services, data collection and analysis, and ethical concerns. Below address some key components addressed in the course it was applied to.

Prompt Engineering

- Beginning lessons typically involve demonstrating how different answers come when questions are worded differently, emphasizing the importance of designing the prompt to fit one’s needs (prompt engineering).
- Students learn about the importance of **context**, informational **queues**, **intent** & audience, and **formatting** within their prompt commands in order for LLMs to deliver the desired output.
- Through prompt engineering for programming code, students are required to “sketch out” the desired code to be produced, essentially pre-coding the steps for the necessary result.
- This idea is analogous to technical report specifications, which allows students to practice explaining tasks proficiently enough so that others understand for procedures to be reproduced.
- Potentially, we may reach an age where “coding” becomes less focused, and pre-coding to an AI engine will be the norm.

Reverse Outlining

- LLMs also are useful for *reverse outlining*, which relay whether the LLM detects the main points and organization of the submission.
- *“What are the main components of this passage?”*
- Students can plug in both their programming code as well as statistical writing to check whether the they were able to clearly relay their message.
- Such techniques can be utilized when tutoring isn’t available.

Hallucination Detection and Quality Control

- Sometimes, LLMs produce hallucinations, generating a response that is either factually incorrect, nonsensical, or disconnected from the input prompt.
- This issue is because the responses are pulled from text data mining and not from actual knowledge.
- Students need to be taught to identify such issues and take advantage It will be up to you to provide clear understanding of the process as to avoid these mishaps.

Product Management

- Through prompt engineering, hallucination detection, and general quality control, we train students to focus instead on the *human* components of research, review, and production.
- As emphasized by Tu et all (2024), we transform students into “product managers rather than software engineers” and demonstrate how to simplify menial tasks and focus more on checking whether the LLM provided adequate results.

Ethical Concerns

- Like all areas of statistics, there are ethical issues that should be addressed, such as environmental impact, privacy, intellectual property, and ownership.
- Think about the current arguments between art that is AI-generated versus human-generated.
- To reinforce the issue of avoiding plagiarism, emphasize the importance of one’s voice to describe the personal narrative.

AI Output Critique:
An Alternative to Peer Review

Background of Situation

- During previous iterations of this course, students would be assigned sections of another’s paper (the *Statistical Background and Terminology*) to submit a peer review and critique the work.
- This process involved setting a lecture to have student read and review each other’s section, with a guided worksheet to assess the impact of their peer’s work and understanding, to be returned to the student at the end of the session.
- The process was performed during Paper 3 of the course, conducted after requiring 1-on-1 conferencing between the student and instructor/TA for Papers 1 and 2.

Problems with Peer Review

- Most students, from qualitative interviews as well as in previous survey questionnaires, expressed their dislike of this assignment and did not believe that it was useful for their understanding of the material.
- Some reasons behind these feelings were because:
  (1) *They did not feel comfortable with their own understanding of the material, so critiquing another’s work felt intimidating and unclear.*
  (2) *They did not like critiquing their fellow learners and would instead provide feedback that was not critical of their work.*
- In effect, the assessment did not appear to enhance student competency or properly measure their ability to critique an outside work.

A Shift to AI Critique

- Instead of evaluating assessment of their peers, students were given an article from a fictional digital content website (*FuzzRead*) to assess the adequacy of its description of the statistical background (and whether it is AI jargon) and the conclusions drawn by the results in the provided procedure.
- As the lessons on artificial intelligence and LLMs discussed issues with terse writing responses as well as hallucinations, students were made aware of why simply pulling from these programs are not only forms of plagiarism, but also often not suitable for audiences of their reports.
- ChatGPT3 was asked to provide a two-paragraph response to the following prompt:
  Explain: Experiment (compared to an observational study), Response Variable, Dependent variables, factors, levels, main effects, interaction, profile plot, ANOVA Tables, hypothesis testing with experimental design.
- Most student reports (1-2 pages) on the critical assessment provided a more-honest critique on the lack of clarity of the writing (compared to their responses of their peers) and were able to pinpoint elements where statistical communication and analysis could be improved within the passage.
- Furthermore, they were given the opportunity to witness how writing from LLMs loses a sense of voice and perspective, as well as how easily it is to detect.

Student Evaluation Comparison

- Students expressed a preference for critiquing a robot’s assignment compared to one of their peers, as they were much more critical of its output.
- In end-of-course surveys, the assessment scored better overall compared to peer review sessions from prior terms.

Table 1: Student perceptions of the benefit of AI critique compared to peer review in respect to paper writing.

Benefit to Final Draft	Very Much Yes	Somewhat Yes	Somewhat No	Very Much No
AI Lesson and Article Critique [1]	7 (22.2%)	15 (50.0%)	5 (16.7%)	3 (11.1%)
Peer Review Sessions [2]	4 (12.5%)	13 (40.6%)	12 (37.5%)	3 (9.4%)

[1] *AI Lesson and Article Critique were incorporated in Spring 2024 (n = 30).*
[2] *Peer Review Sessions were incorporated in Spring 2023 (n = 32).*

Suggestions for Improvement

- While students found the time spent on these lessons and assessments to be more beneficial compared to peer review sessions, some students did not find it exceptionally helpful, especially in comparison to one-on-one conferencing.
- Furthermore, instruction on AI and assessment was limited to a couple lectures near the end of the course, after much time was spent learning programming fundamentals.
- As one tries to juggle teaching statistical concepts, coding, writing, and AI, the material feels a bit cramped and time-constrained.
- Furthermore, as AI development expands further, these techniques will expand, and we will likely shift again how we implement
- Modified formats could include
  (1) increase the number of credits for the class (and increase lecture time).
  (2) create an additional mini-course on AI for students to take either concurrently or after their programming requirement.
  (3) cover the material over a two-course sequence with added topics (or keep the amount for a quarter system

References

- Tu, X., Zou, J. Su, W., Zhang, L. (2024). What should data science education do with large language models? Harvard Data Science Review, 6(1), 1–28.
- Wang, S., Wang, F. Zhu, Z., Wang, J., Tran, T., Du, Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252, 1–19.

Acknowledgements

- GWU Writing in the Disciplines Program.
- GWU Department of Statistics.
- UCLA Department of Statistics and Data Science.
- USCOTS 2025 Posters & Beyond Co-Chairs (for their suggestions and guidance).