<https://github.com/jdonohue44/Simulated-RCT-Generative-AI-Can-Harm-Learning/tree/main>

For Wednesday, 12/4 @5pm Uris:

1. Create data - Audrey
2. Start writing code - Jared/Audrey
3. Visualizations/Plots/Table - Jared/Audrey

Can also meet again Sunday 12/8 or Monday 12/9.

Project is due at midnight 12/9.

## High-level Instructions

Throughout this semester, we will cover various topics related to online experiment design and analysis, ranging from remedies and designs to solve SUTVA violations to experiment diagnostics (e.g. SRM and simulated AA) to ML-based estimation techniques. In this project, you will be implementing an algorithm or a method described in a conference paper (e.g. KDD) or an Engineering blog post.

### Exec Summary

### Context

* **What problem are you solving? (problem description)**
  + As AI-based chatbots have become more widely used, students have naturally began relying on them for homework, studying, and essay writing. However, these chatbots have not been around very long, so the effects on student’s lives long term is not well understood. Many discussions around the use of AI focus on increasing worker productivity and how to keep students from cheating with AI, rather than analyzing the impact that AI use has on learning new skills.
  + This paper aims to quantify the impact that use of AI has on students' performance. It does so by creating two distinct chatbots: one modeled after GPT-4 and another designed specifically to promote active learning. Student performance is then evaluated both with and without access to these chatbots.
* **What is the practical application or significance of it? (Why it is worth solving)**
  + While AI tools like GPT-4 have shown significant potential to enhance productivity and provide knowledge, they also pose risks, such as inhibiting learning through overreliance or reducing the development of foundational skills. Understanding this tradeoff is essential for ensuring that AI technologies are used responsibly and effectively, particularly in contexts like education where long-term skill development is crucial.
  + There has been an increase in students using AI, and current lesson plans were not designed with that in mind. It would be difficult, at this point, to fully eliminate the use of AI by students. For that reason, it is important to understand the effect that it has on learning. This research represents a first step toward integrating AI into education effectively. For example, if the chatbot designed to promote learning turns out to be beneficial, or at the very least not harmful, it could be a replacement for student use. Or, if we understand the impact on learning, teachers may be able to re-design lesson plans to let students use AI in a way that still promotes learning and adoption of new skills.
* **Describe in detail the design and method in the paper using your own language.**
  + The study examined the educational impact of two AI chatbots—GPT Base and GPT Tutor—on student learning outcomes. GPT Base functioned as a standard GPT-4 assistant, delivering typical answers to prompts and acting as a straightforward tutor. GPT Tutor, on the other hand, was designed with additional features to support deeper learning. It provided multiple correct answers, incorporated teacher feedback to address common mistakes, and offered hints rather than full answers to encourage active problem-solving.
  + The research design was a randomized controlled trial involving nearly 1,000 students from 50 classrooms in grades 9 through 11 at a large high school in Turkey. Randomization was performed at the classroom level because students were already randomly assigned to these groups, and honors classrooms were excluded because they are not randomly assigned. The study spanned four 90-minute sessions, each with a sequence of 3 activities. First, teachers reviewed the topic with students. Second, students participated in a randomized, assisted AI session, where they either used GPT Base, GPT Tutor, or relied on textbooks and notes (control group). This session was scored to assess performance. Finally, students completed an unassisted evaluation, also scored, to measure how well they had learned the new information.
  + Data was collected in three main ways. At the start of the study, students completed a survey capturing their demographics and educational background. During the sessions, student performance was recorded for both the assisted practice and the unassisted evaluations. Additionally, students who interacted with AI chatbots had their chat data logged, and surveys captured their experiences using the tools.
  + To evaluate the impact of the interventions, the authors used a regression model to analyze student outcomes. The dependent variable, *Outcome(j) ,* represented the normalized grade of a student in either the assisted (j = 0) or unassisted (j = 1) sessions, scaled from 0 to 1.
  + The independent variables *GPTBasec* and *GPTTutorc* indicate the treatment group for each class. The model controlled for prior student performance using normalized GPA from the previous year, PrevGPAi, and included fixed effects for session, grade, year, and time-related variations (θs, δg, αy, λt). Errors were clustered at the classroom level to account for correlations within groups.
  + Results indicate that use of the chatbots increased performance on the assisted assessment, with GPT Base improving scores by .137 (out of 1) and GPT Tutor improving scores by .361 (out of 1) relative to the control group. On the unassisted assessment, GPT Base decreased performance by .054 (out of one) relative to the control group (17% decrease). GPT Tutor’s impact on the unassisted portion was statistically significant at -.004.
  + Stopped right before problem level specification

## Methodology Review

* **Replicate the method or algorithm proposed by the paper/blog post using simulated or fake data to test the algorithm. You are encouraged to use your own dataset as well.**
  + Data: 1,000 students, each with:
    - Classroom (1-50)
    - Assisted Assessment Normalized Grade (0-1)
    - Unassisted Assessment Normalized Grade (0-1)
    - Group Assignment (Randomly Assigned through Classroom unit)
  + Algorithm:
    - 0/ Pre-Experiment A/A to check for existing differences/variance
      * Data was collected in three main ways. At the start of the study, students completed a survey capturing their demographics and educational background. During the sessions, student performance was recorded for both the assisted practice and the unassisted evaluations. Additionally, students who interacted with AI chatbots had their chat data logged, and surveys captured their experiences using the tools.
    - 1/ Randomly assign classrooms (C, T1, or T2)
      * C: no GPT, textbooks and notes only
      * T1: GPT-Base
      * T2: GPT-Tutor
    - 2/ Simulate running the assisted and unassisted assessments
    - 3/ Calculate scores for each group
      * To evaluate the impact of the interventions, the authors used a regression model to analyze student outcomes. The dependent variable, *Outcome(j) ,* represented the normalized grade of a student in either the assisted (j = 0) or unassisted (j = 1) sessions, scaled from 0 to 1.
      * The independent variables *GPTBasec* and *GPTTutorc* indicate the treatment group for each class. The model controlled for prior student performance using normalized GPA from the previous year, PrevGPAi, and included fixed effects for session, grade, year, and time-related variations (θs, δg, αy, λt). Errors were clustered at the classroom level to account for correlations within groups.
    - 4/ Compare scores using statistical test and interpret results
      * Results indicate that use of the chatbots increased performance on the assisted assessment, with GPT Base improving scores by .137 (out of 1) and GPT Tutor improving scores by .361 (out of 1) relative to the control group. On the unassisted assessment, GPT Base decreased performance by .054 (out of one) relative to the control group (17% decrease). GPT Tutor’s impact on the unassisted portion was statistically significant at -.004.

## Conclusion

## Citation

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486>

## 

## Potential Papers

**AI**

1. **✅**[Generative AI Can Harm Learning](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486) **[CHOSEN]**

**Health Care**

1. ​​[Applying A/B Testing to Clinical Decision Support: Rapid Randomized Controlled Trials](https://pmc.ncbi.nlm.nih.gov/articles/PMC8065554/)
   1. This one goes through different alerts for Clinical Decision Support (CDS) that pop up in an EHR and encourage the doctor or nurse to take a specific action. The goal is to make the alert as clear as possible so that the doctor or nurse acts on it when it is first seen, instead of dismissing it and having it pop up again later.
   2. We could try to simulate data for different versions of the alert that are in accordance with the results they see in the paper, and then run analysis on that data.
2. [Effectiveness of a Behavior Change Technique–Based Smartphone Game to Improve Intrinsic Motivation and Physical Activity Adherence in Patients With Type 2 Diabetes: Randomized Controlled Trial](https://games.jmir.org/2019/1/e11444/)

**Behavioral Economics**

* [CELSS Lab](https://celss.iserp.columbia.edu/) - Columbia Experimental Laboratory for Social Sciences

**Business/Product**

1. [The Consumer Welfare Effects of Online Ads: Evidence from a 9-Year Experiment](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4877025)
2. [[2311.14698] Business Policy Experiments using Fractional Factorial Designs: Consumer Retention on DoorDash](https://arxiv.org/abs/2311.14698)
3. [How Zocdoc Uses Data to Make Better Products for Patients](https://zocdoc-engineering.medium.com/how-zocdoc-uses-data-to-make-better-products-for-patients-b349f5b3ffdf)
4. [Is There Such a Thing as a Bad Ad? | Columbia Business School](https://business.columbia.edu/research-brief/online-marketplaces-retail-media)
5. [Onsite Retargeting: A Randomized Field Experiment](https://repub.eur.nl/pub/104608/)
6. [Why Every Runner Should Use WHOOP | Project PR](https://www.whoop.com/us/en/thelocker/project-pr-runner-study/)