Jared Donohue

Pre-doctoral Research Associate
at Columbia Business School
Data Science Institute Scholar

About Me

Current Research at Columbia Business School

- Running online RCT on AI-assisted coding environment behavior (Prof. David Holtz)
- Diff-in-diff analysis of firm-level AI agent adoption for experimentation (Prof. David Holtz)
- Leading an LLM-assisted meta-analysis of behavioral clinical trial compliance (Prof. Hannah Li)

Methods & Training

- M.S. in Data Science with focus on causal inference
- B.S. in Computer Science with lots of hands-on programming experience (R, Python, Java, C)
- Published in *Scientific Data*: open, reproducible LLM-assisted extraction pipeline
- 6+ years doing analysis, engineering, and running experimentation at Amazon

Motivation

- Methodologically interested in causal inference and practically driven by real-world impact (e.g. solving the problems of AI in education)

Does Generative AI Use Harm Learning? A Simulated Reproduction and Extended Analysis

- Simulated data from: *Generative AI without guardrails can harm learning: Evidence from high school mathematics* (Bastani et al, 2024) to validate statistical reproducibility and understand causal analysis pnas.org/doi/10.1073/pnas.2422633122
- Analyzed author-shared data to investigate covariates and AI usage
- Rigorous causal inference in educational setting
- Open science and reproducibility
- *github.com/jdonohue44/Simulated-RCT-Generative-AI-Can-Harm-Learning*

Bastani et al. (2024) Experiment Design

Pre-registered at https://aspredicted.org/4DL_Q3J

- **Setting**: 50 classrooms of 943 high school students in Ankara Turkey, Fall 2023
- **Design**: 3-arm RCT, randomized by classroom
  - Control (no AI access) – 19 classrooms
  - GPT-4-Base (raw chatbot) – 15 classrooms
  - GPT-4-Tutor (designed to teach without giving solution) – 16 classrooms
- **Intervention**: Provide variations of AI access during 30-min problem sets
- **Outcome**: Performance on problem sets and subsequent unassisted exam
- **Key Finding**: AI improved assisted performance but reduced unassisted exam scores when AI guardrails (AI Tutor) were not used.
- **Mechanism**: Students use AI as a “crutch” for solutions and not understanding, identified by comparing Base vs Tutor GPT, controlling for incorrect GPT-Base solutions

Simulation

Approach:

- Simulated classroom & student performance data by matching study parameters (e.g. 50 classrooms, 3 experimental arms, 1000 students, exam scores)
- Treatment effect sizes matched in data generating process

Why simulate?

- Understand the experiment design and causal mechanisms
- Validate statistical analysis procedures
- Original data had not been shared
- Opportunity through Design and Analysis of Online Experiments course

Result:

- Successfully simulated treatment effects from paper, that AI improved assisted performance but reduced unassisted exam scores when AI guardrails (AI Tutor) were not used.

Simulated Effect on Exam Performance

AI tools which provide immediate solutions function as a crutch to improve immediate performance but impede durable learning

Problem Sets (Assisted)

Mean Score by Group

Control: 28%
GPT-4-Base: 42% (+48% v Control)
GPT-4-Tutor: 66% (+127% v Control)

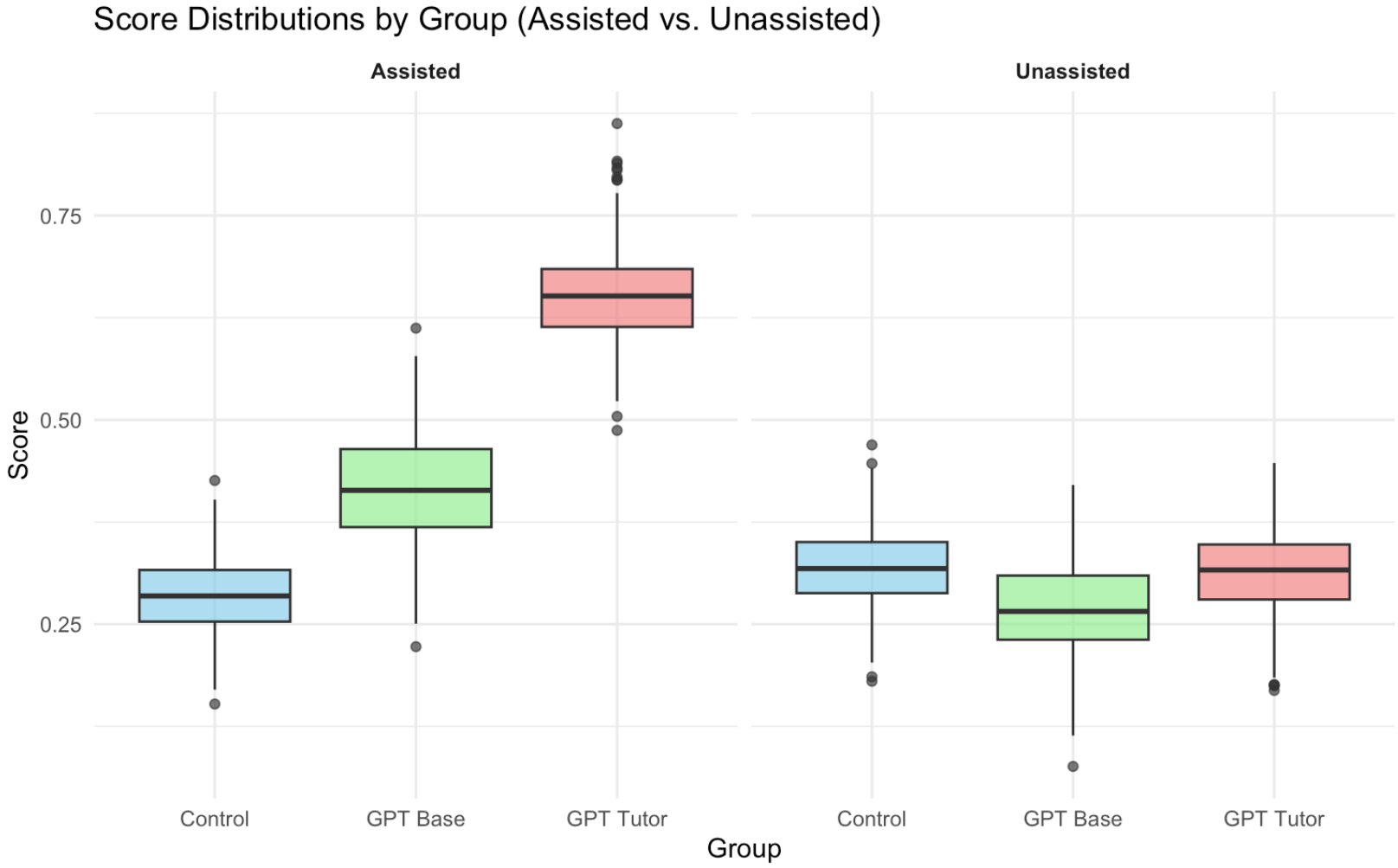
AI groups perform significantly better with tool access

Exam (Unassisted)

Mean Score by Group

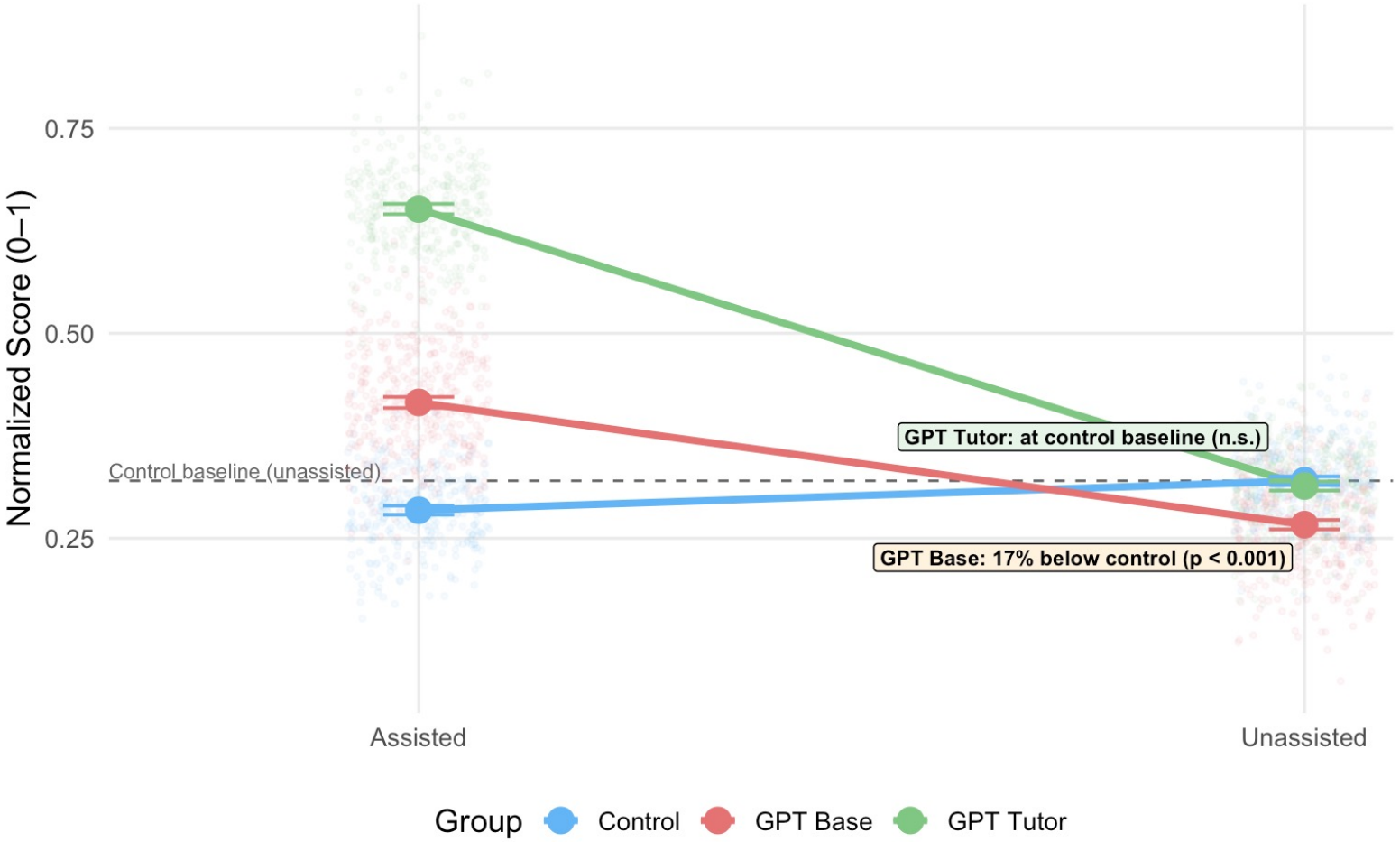
Control: 32%
GPT-4-Base: 27% (-17% v Control)
GPT-4-Tutor: 31% (n.s.)

Without AI, treated groups perform worse


Score Distributions by Group (Assisted vs. Unassisted)
Assisted
Unassisted
Score
0.75
0.50
0.25
Control
GPT Base
GPT Tutor
Control
GPT Base
GPT Tutor
Group

The Crutch Effect: AI Boosts Assisted Scores but GPT Base Harms Unassisted Learning

Mean scores with 95% CI; dashed line = Control unassisted baseline



Extending Analysis with Author-shared Data

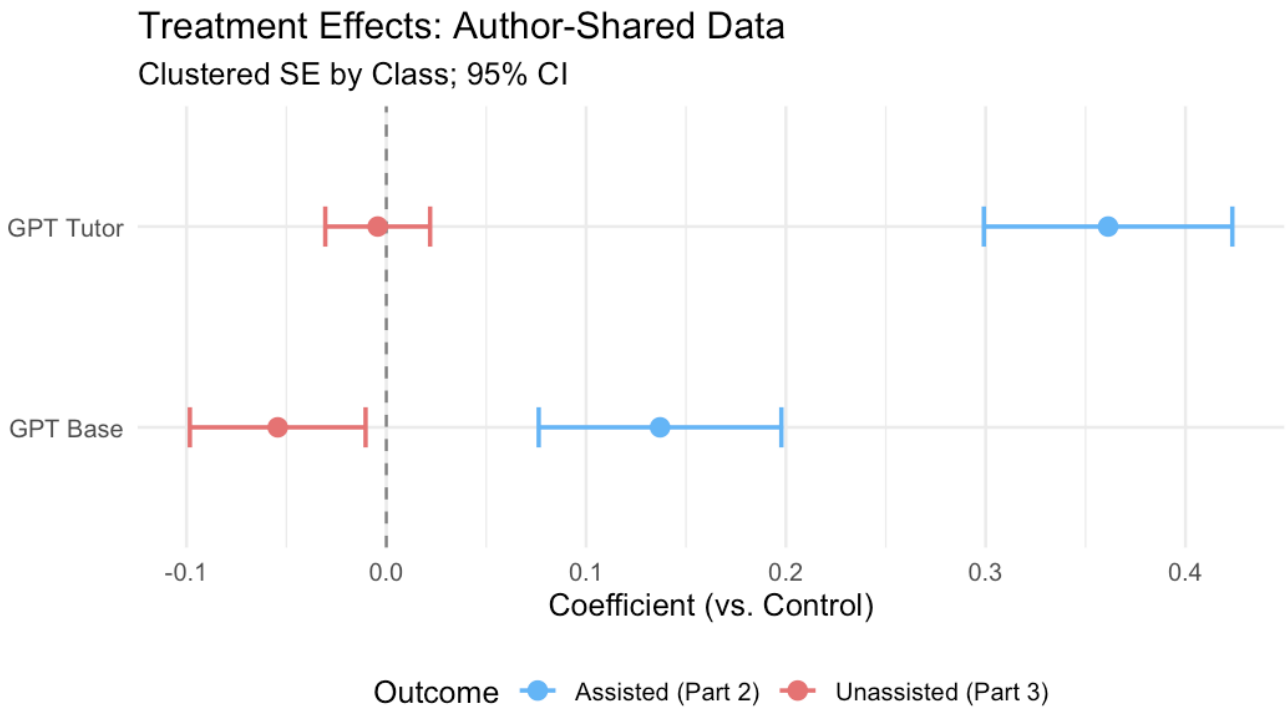
Authors released study data on GitHub in June 2025: github.com/obastani/GenAICanHarmLearning

Author's GitHub now includes:

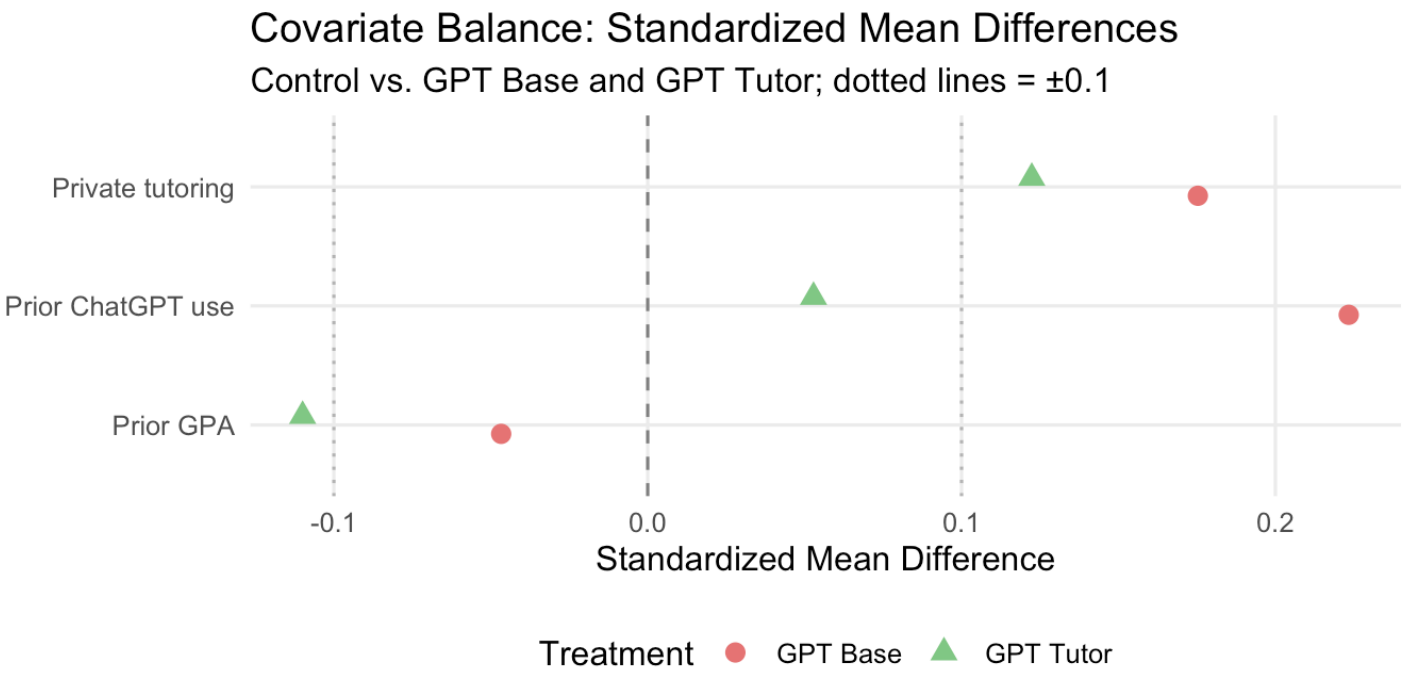
- Full student-level data
- Chat logs (student-AI interactions)
- Student perception surveys
- Problem-by-problem performance

Extended Analysis:

- **Evaluate balance of covariates**, especially important since the schools randomly assigned classrooms, not the researchers
- **Reproduce core regressions** using actual data
- **Confirmed exact number of student participants (n=943):** paper used "nearly a thousand"
- **Heterogeneous effects:** Investigated if students with prior AI experience perform differently



Outcome	Coefficient	Estimate	SE	95% CI Lower	95% CI Upper	p-value	
Assisted (Part 2)	GPT Base	0.137	0.031	0.076	0.198	0.000	***
Assisted (Part 2)	GPT Tutor	0.361	0.032	0.299	0.423	0.000	***
Assisted (Part 2)	Prior GPA	0.802	0.076	0.653	0.952	0.000	***
Unassisted (Part 3)	GPT Base	-0.054	0.022	-0.098	-0.010	0.020	*
Unassisted (Part 3)	GPT Tutor	-0.004	0.013	-0.031	0.022	0.749	n.s.
Unassisted (Part 3)	Prior GPA	1.334	0.069	1.199	1.470	0.000	***



	Variable	Control Mean	GPT Base Mean	GPT Tutor Mean	SMD (Base vs. Control)	SMD (Tutor vs. Control)
gpa_prev	Prior GPA	0.822	0.817	0.810	-0.047	-0.110
private_tutorship	Private tutoring	0.572	0.657	0.632	0.175	0.122
chatgpt_use	Prior ChatGPT use	1.172	1.455	1.238	0.223	0.053

Future Research Directions

- **Measurement**. Do existing evaluation metrics (e.g. exam scores) capture durable learnings important in today's society? What is missing? When should assisted performance be measured vs unassisted performance?

- **Quasi-Experimental Identification**. Can AI policy adoption across Irish/European universities serve as a natural experiment at scale? Leverage causal inference methods:
  - Instrumental Variables
  - Difference-in-Differences
  - Regression Discontinuity Designs

Future Research Direction

- **Individualized Effects**. Which students benefit from AI and which are harmed? Average effects might mask more granular evidence. Consider within-student crossover designs (N-of-1) and always assume heterogeneous treatment effects

- **AI Tool Design for Higher Education**. How should AI tools be designed to reduce negative effects on unassisted learning? E.g., tutor variants that force reasoning through concepts before revealing solutions (or never revealing solutions)

Thank You

Questions?

Jared Donohue
jjd2203@columbia.edu

github.com/jdonohue44/Simulated-RCT-Generative-AI-Can-Harm-Learning