Exploring Bias and Creativity: A Comparative Study of Human and AI Generated Works with AI Priming*

Final Project - DATASCI 241

Angela Chang, Dean Nakada,
Kevin Stallone, Dave Zack

*Generated by ChatGPT 3.5

Research question

When primed, are people more likely to be convinced something is written by a large language model, AI, such as ChatGPT?

Hypothesis

Null Hypothesis (H_0): There is no significant difference between control and treatment groups when it comes to the rate of selecting A.I. as the passage author. As a result, there will be no significant difference in accuracy rates between control and treatment groups.

Alternative Hypothesis (H_a): Exposure to an article about the recent advancements in the capabilities of A.I., the treatment, will lead people to believe that passages they read are authored by A.I. As a result, for the treatment group, accuracy on AI-generated pieces will increase, but accuracy on human-written pieces will decrease.

Treatment

Treatment - Generative AI

AI-generated text, from tools like ChatGPT, is starting to impact daily life. Teachers are testing it out as part of classroom lessons. Marketers are champing at the bit to replace their interns. Memers are going buck wild. Me? It would be a lie to say I'm not a little anxious about the robots coming for my writing gig. (ChatGPT, luckily, can't hop on Zoom calls and conduct interviews just yet.)

With generative AI tools now publicly accessible, you'll likely encounter more synthetic content while surfing the web. Chances are you have already interacted with a large language model if you've ever used an application — like Gmail — that includes an autocomplete feature, gently prompting you with the word ‘‘attend’’ after you type the sentence ‘‘Sadly I won't be able to….’’ But autocomplete is only the most rudimentary expression of what software like GPT is capable of. It turns out that with enough training data and sufficiently deep neural nets, large language models can display remarkable skill if you ask them not just to fill in the missing word, but also to continue on writing whole paragraphs in the style of the initial prompt. Algorithms with the ability to mimic the patterns of natural writing have been around for a few more years than you might realize.

Placebo - Plant-based Meat

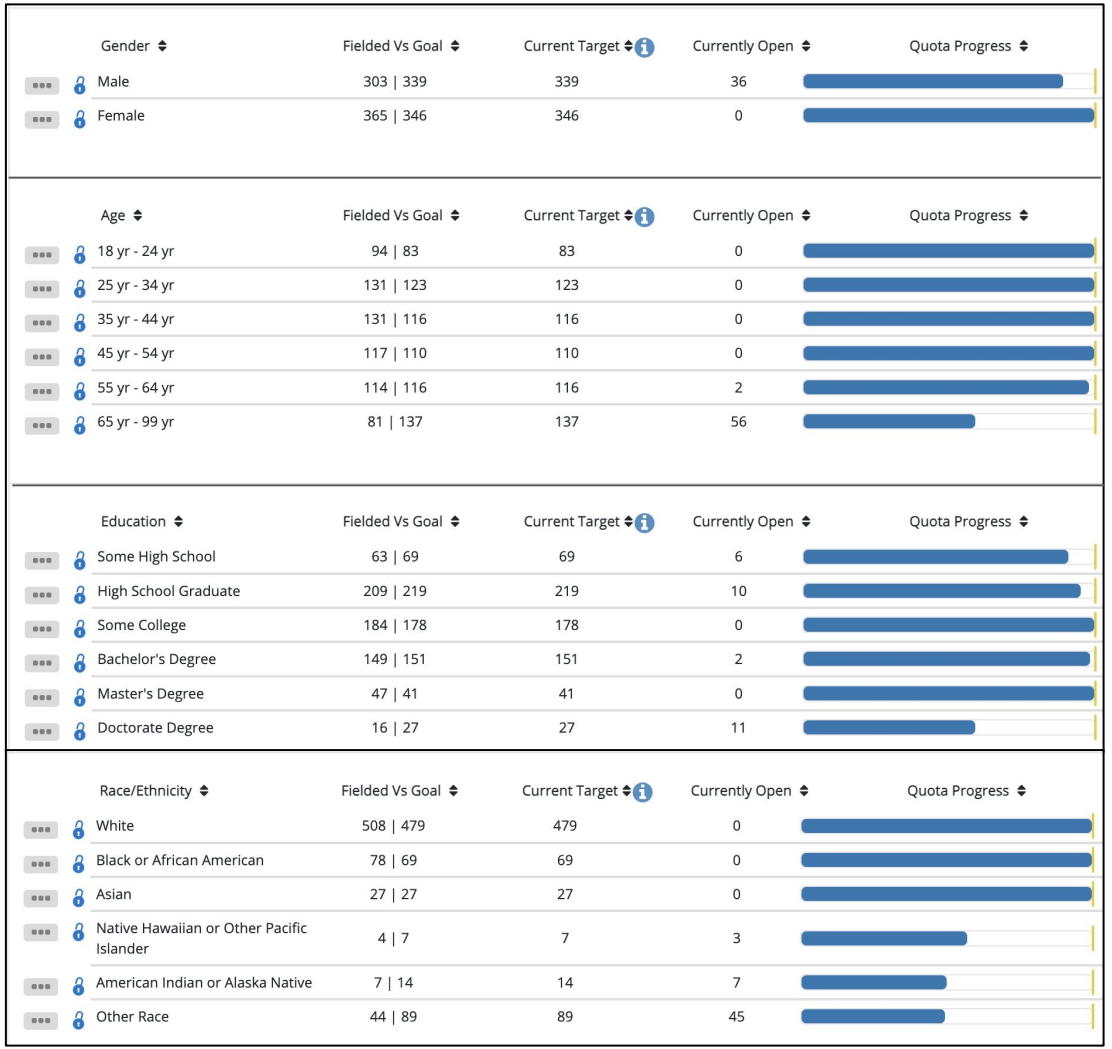
Plant-based meat has increasingly been making headlines in recent years, as product sales swell and industry investment continues to reach dizzying heights. "Plant-Based Meat 2021-2031", a new report by IDTechEx, explores the technologies that are shaping the plant-based meat industry, alongside the consumer and market factors that will decide whether plant-based meat can truly disrupt the $1 trillion global meat industry.

Modern plant-based meat companies such as Beyond Meat and Impossible Foods have invested heavily into R&D and technology development to make their products as realistic as possible. Impossible Foods uses a genetically engineered strain of yeast to produce soy leghemoglobin, a key ingredient that makes its meat substitutes "bleed" and gives them a uniquely meaty flavor. To create its Beyond Burger product, Beyond Meat uses a food extrusion machine originally developed at the University of Missouri, which uses heat and pressure to force plant proteins into a fibrous, meat-like texture that resembles muscle fibers. Rather than using genetic engineering to produce its products, the Beyond Meat burger uses beet juice to replicate the bleeding from a real burger. Coconut oil and cocoa butter are used to provide marbling to further replicate the texture of real meat.

This technology development has helped to create a new generation of plant-based meat products that are slowly winning over meat-eating consumers. Now, companies across the world are working to leverage a range of emerging technologies to help create the next generation of plant-based products.

Measurement units

- Hosted on PureSpectrum which linked to the Berkeley Qualtrics survey we designed
- Used PureSpectrum's built in quota system to get a wide array of people to take the survey
- Overall there were a total of 668 respondents who completed the survey during the PureSpectrum fielding
- Measured people's responses to the various questions listed in the survey

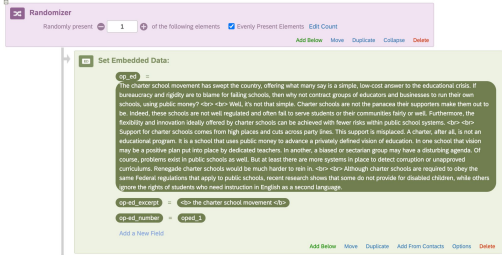


Gender	Fielded Vs Goal	Current Target	Currently Open	Quota Progress
Male	303 | 339	339	36	
Female	365 | 346	346	0	

Age	Fielded Vs Goal	Current Target	Currently Open	Quota Progress
18 yr - 24 yr	94 | 83	83	0	
25 yr - 34 yr	131 | 123	123	0	
35 yr - 44 yr	131 | 116	116	0	
45 yr - 54 yr	117 | 110	110	0	
55 yr - 64 yr	114 | 116	116	2	
65 yr - 99 yr	81 | 137	137	56	

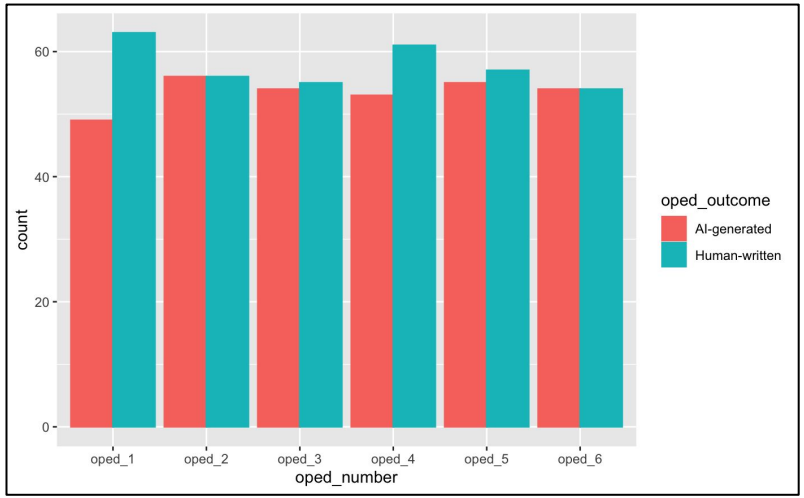
Education	Fielded Vs Goal	Current Target	Currently Open	Quota Progress
Some High School	63 | 69	69	6	
High School Graduate	209 | 219	219	10	
Some College	184 | 178	178	0	
Bachelor's Degree	149 | 151	151	2	
Master's Degree	47 | 41	41	0	
Doctorate Degree	16 | 27	27	11	

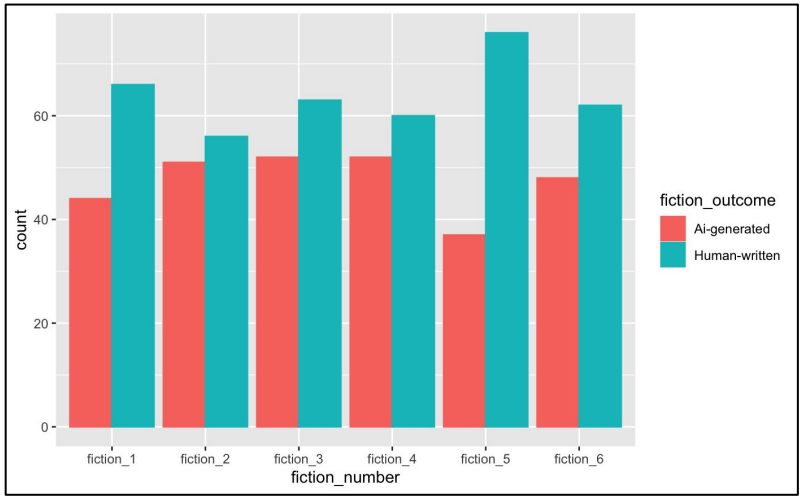
Race/Ethnicity	Fielded Vs Goal	Current Target	Currently Open	Quota Progress
White	508 | 479	479	0	
Black or African American	78 | 69	69	0	
Asian	27 | 27	27	0	
Native Hawaiian or Other Pacific Islander	4 | 7	7	3	
American Indian or Alaska Native	7 | 14	14	7	
Other Race	44 | 89	89	45	



Randomization

- Multiple levels of randomization in the study all handled by built-in Qualtrics functionality
  1. Control/treatment assignment
  2. Oped piece assignment (six total pieces - three human-written and three AI generated)
  3. Fiction piece assignment (six total pieces - three human-written and three AI generated)
  4. Demographic question order was randomly displayed
  5. Outcome check was displayed in a random order

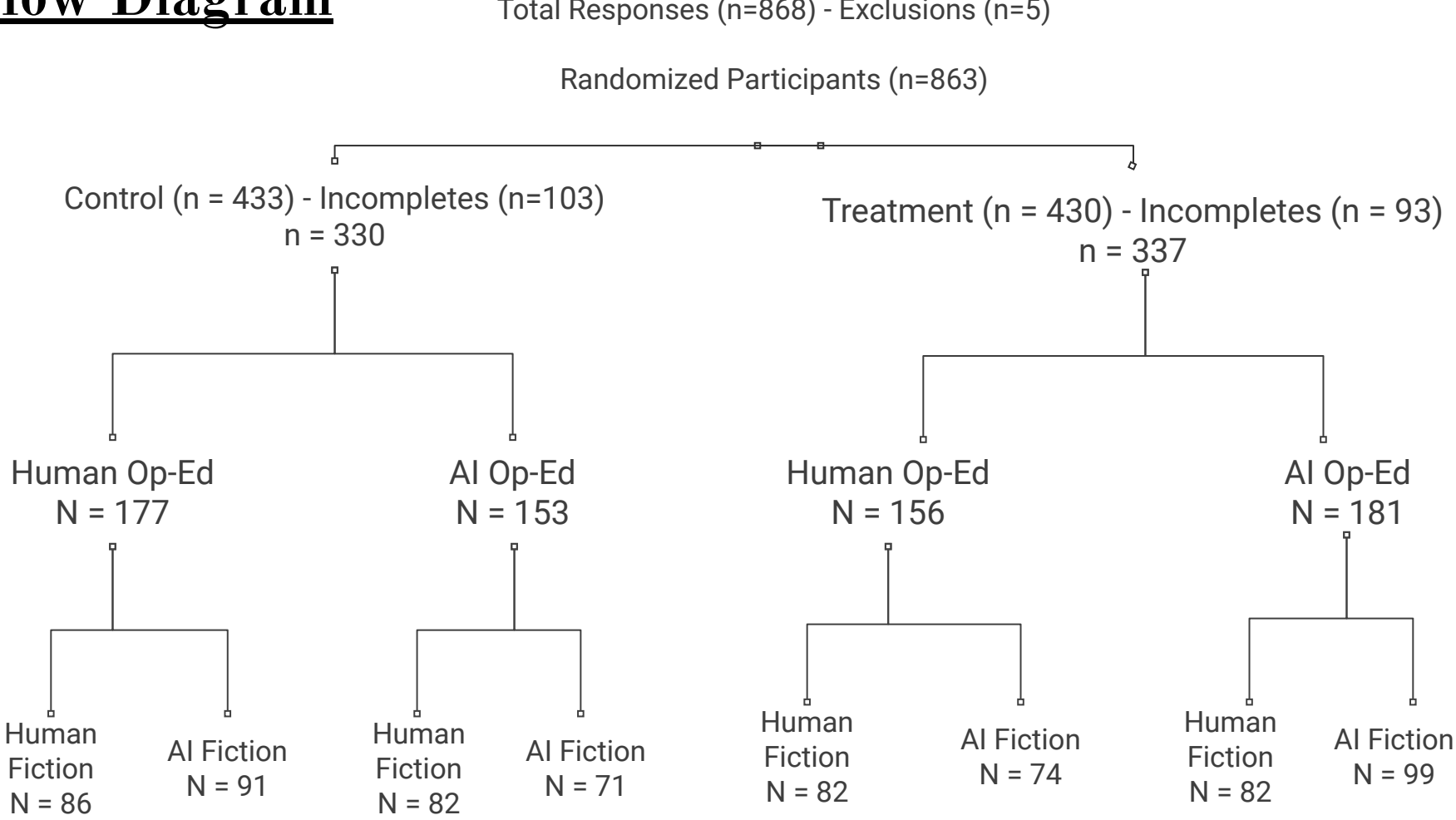




ROXO for the 2 x 2 x 2 Factorial Design

Group	Treatment	Op-Ed	Fiction	Outcome 1	Outcome 2
R	X	X_1	X_1	O	O
R		X_1	X_1	O	O
R	X	X_2	X_2	O	O
R		X_2	X_2	O	O
R	X	X_1	X_2	O	O
R		X_1	X_2	O	O
R	X	X_2	X_1	O	O
R		X_2	X_1	O	O

Flow Diagram



Passages

1. Op-Ed: **“Charter Schools Won’t Save Education” by Gary Orfield, 1998**
2. Op-Ed 2: “American Health Care Is Broken. Major Hospitals Need to Be Part of the Solution” by Donald M. Berwick and Michelle A. Williams, 2023
3. Op-Ed 3: **“Opinion: Air quality disasters aren’t going away. Here’s how we should protect ourselves”** by Abraar Karan, Devabhaktuni Srikrishna, and Ranu Dhillon, 2023
4. Fiction 1 (novella): “Hunger” by Lan Samantha Chang
5. Fiction 2 (short story): “The Visitor” by Lydia Davis
6. Fiction 3 (novel): Stranger by Sherwood Smith and Rachel Manija Brown

Example Passages: Fiction

Human-Written Text "Hunger" by Lan Samantha Chang	AI-Generated Text ChatGPT
"Ruth grew tall and almost painfully slender; in the sixth grade, she began to practice on Tian's full-sized violin. At fourteen, she won a city-wide competition for young musicians. As the winner, she would perform a concerto, accompanied by the orchestra that had sponsored the competition. Tian had chosen the Sibelius concerto, and for many, many months he and Ruth practiced those wondrous loops of sound, the joyous slithering movements of Sibelius."	"Amelia grew up in the shadow of her parents' brilliance. Her father was a renowned surgeon, her mother, a celebrated physicist, and her older siblings, champions in their respective fields. All of them stood tall, their voices firm and resounding. In their world of assertive conversations and challenging ideas, Amelia was a quiet echo."

Outcome Measures

Do you think the passage about [*piece assigned]* was:

- AI-Generated (random order)
- Human-Written (random order)

1. Percent of subjects who answer that the op-ed was AI-generated
2. Percent of subjects who answer that the fiction piece was AI-generated
3. Op-ed writer accuracy (human or AI)
4. Fiction writer accuracy (human or AI)

Results and Analysis: estimated ATE on “AI rate”

Estimated ATE on AI rate

	AI Rate	
	Op-ed	Fiction
	(1)	(2)
Treated (Priming)	0.101***	0.123***
	(0.039)	(0.038)
Constant	0.430***	0.364***
	(0.027)	(0.027)
Observations	667	667
R^2	0.010	0.015
Adjusted R^2	0.009	0.014
Residual Std. Error (df = 665)	0.498	0.491
F Statistic (df = 1; 665)	6.843***	10.450***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Focus of analysis

- Outcome: % of subjects perceiving a passage as generated by AI (“AI rate”)

Result

- **Significantly more people selected AI as an author** for both op-ed and fiction passages.

Hypothesis

Null Hypothesis (H_0): There is no significant difference between control and treatment groups when it comes to the rate of selecting A.I. as the passage author. As a result, there will be no significant difference in accuracy rates between control and treatment groups.

Results and Analysis: estimated ATE on “accuracy rate”

Estimated ATE on accuracy

	Accuracy	
	Op-ed	Fiction
	(1)	(2)
Treated (Priming)	0.055	-0.028
	(0.039)	(0.039)
Constant	0.476***	0.497***
	(0.028)	(0.028)
Observations	667	667
R^2	0.003	0.001
Adjusted R^2	0.002	-0.001
Residual Std. Error (df = 665)	0.500	0.500
F Statistic (df = 1; 665)	2.047	0.527
Note:	*p<0.1; **p<0.05; ***p<0.01	

Focus of analysis

- Outcome: % of accurate answers about the author of a passage (“accuracy rate”)

Result

- There was **no significant treatment effect on the accuracy** for both op-ed and fiction passages.

Hypothesis

Null Hypothesis (H_0): There is no significant difference between control and treatment groups when it comes to the rate of selecting A.I. as the passage author. As a result, there will be no significant difference in accuracy rates between control and treatment groups.

This cannot be rejected.

Heterogeneous effect on accuracy (op-ed)

Interactive estimated ATE on accuracy

	Accuracy			
	Passage type: Op-ed		Passage type: Fiction	
	(1)	(2)	(3)	(4)
Treated (Priming)	0.058	-0.042	-0.025	-0.149***
	(0.039)	(0.055)	(0.038)	(0.054)
Op-ed type: AI	-0.042	-0.144***		
	(0.039)	(0.055)		
Treated * Op-ed type: AI		0.202***		
		(0.077)		
Fiction type: AI			-0.148***	-0.273***
			(0.038)	(0.053)
Treated * Fiction type: AI				0.248***
				(0.076)
Constant	0.495***	0.542***	0.569***	0.631***
	(0.033)	(0.038)	(0.033)	(0.037)
Observations	667	667	667	667
R^2	0.005	0.015	0.023	0.038
Adjusted R^2	0.002	0.010	0.020	0.034
Residual Std. Error	0.500 (df = 664)	0.498 (df = 663)	0.495 (df = 664)	0.492 (df = 663)
F Statistic	1.603 (df = 2; 664)	3.347** (df = 3; 663)	7.686*** (df = 2; 664)	8.729*** (df = 3; 663)
Note:				*p<0.1; **p<0.05; ***p<0.01

Result

- F-test for model (1) and (2) shows $p < 0.01$.
- **Reject the null hypothesis that the ATE on accuracy remains constant** between AI-generated op-eds and human-written op-eds.
- In-depth analysis revealed that **the treatment drove accuracy for AI-generated op-eds** significantly.
- No significant effect was seen for human-written op-eds.

Heterogeneous effect on accuracy (fiction)

Interactive estimated ATE on accuracy

	Accuracy			
	Passage type: Op-ed		Passage type: Fiction	
	(1)	(2)	(3)	(4)
Treated (Priming)	0.058	-0.042	-0.025	-0.149***
	(0.039)	(0.055)	(0.038)	(0.054)
Op-ed type: AI	-0.042	-0.144***		
	(0.039)	(0.055)		
Treated * Op-ed type: AI		0.202***		
		(0.077)		
Fiction type: AI			-0.148***	-0.273***
			(0.038)	(0.053)
Treated * Fiction type: AI				0.248***
				(0.076)
Constant	0.495***	0.542***	0.569***	0.631***
	(0.033)	(0.038)	(0.033)	(0.037)
Observations	667	667	667	667
R^2	0.005	0.015	0.023	0.038
Adjusted R^2	0.002	0.010	0.020	0.034
Residual Std. Error	0.500 (df = 664)	0.498 (df = 663)	0.495 (df = 664)	0.492 (df = 663)
F Statistic	1.603 (df = 2; 664)	3.347** (df = 3; 663)	7.686*** (df = 2; 664)	8.729*** (df = 3; 663)
Note:				*p<0.1; **p<0.05; ***p<0.01

Result

- F-test for model (3) and (4) shows $p < 0.01$.
- **reject the null hypothesis that the ATE of priming on accuracy remains constant** between AI-generated fiction and human-written fiction.
- In-depth analysis revealed that **the treatment decreased accuracy for human-written fictions significantly.**
- No significant effect was seen for AI-generated fictions.

Hypothesis

Null Hypothesis (H_0): There is no significant difference between control and treatment groups when it comes to the rate of selecting A.I. as the passage author. As a result, there will be no significant difference in accuracy rates between control and treatment groups.

Heterogeneous treatment effect conditional on the author of a passage was found!

Summary of the result

- Priming is likely to affect the probability people perceive an author of a passage as AI.
- Priming is unlikely to affect people's accuracy about the author
- There is a potential heterogeneous ATE on accuracy, conditional on the author of passage.
    - Subjects' accuracy on AI-generated pieces increased as a result of priming. The impact was significant for op-ed.
    - Subjects' accuracy on human-written pieces decreased as a result of priming. The impact was significant for fiction.

Questions & Concerns

- Sample size for the covariate buckets
- Only one language model used, ChatGPT
- Testing subject compliance further - did people actually read the passages?