Zachary Robers

Mr. Nemitz

AP Statistics (7)

May 25, 2023

Quality and Recognizability of ChatGPT Responses in Comparison to Human-Authored Scholarship

I. Introduction

ChatGPT, a generative AI software created by OpenAI, is disrupting classrooms across the United States. The software, which uses a transformer neural network to generate responses to various prompts and questions, is being utilized by students to write essays and complete assignments. Many schools and nations have responded to the rise of ChatGPT by administering a ban on its use. However, only about half of students (51%) consider the use of the AI model to be a form of cheating. Since inputting essay prompts and academic questions into ChatGPT enables students to avoid putting in the effort to develop academically, it is easy to see how the software can negatively affect the United States education system. Perhaps most paramount to preventing the use of ChatGPT is ensuring that teachers can detect its use. When teachers can detect the use of ChatGPT, they can administer proper consequences, thus discouraging its future use. Since ChatGPT was pre-trained by an abundance of written works and designed to mimic human writing styles, detecting when it has been used is difficult, especially when students take precautions to make their use of ChatGPT less apparent. The following study aims to determine how effective high school teachers are at detecting when ChatGPT has been used to answer an

academic question. Furthermore, the study analyzes teachers' perceptions of the quality of responses written by ChatGPT in comparison to human-authored responses.

II. Data Collection

The intended population for inference in this study is all high school teachers in the United States. The teachers at East Chapel Hill High School, a public high school in Chapel Hill, North Carolina with approximately 1400 students, are treated as a cluster sample and the teaching body at the school is assumed to be representative of the teaching body at any high school across the U.S.

The 90 teachers currently employed at East Chapel Hill High School were sent an email asking them to participate in a study to determine the recognizability of ChatGPT responses in comparison to human-authored scholarship. In the email was a link which randomly redirects those who click on it to one of twenty forms.

These twenty forms are divided into four categories representing different subject areas associated with academic questions. The four categories are philosophy, sciences, history, and writing/literature. Each segment contains five forms. All twenty forms follow the same structure, and include the guidelines for participation, the academic question, and the two responses (one generated by ChatGPT and a human-authored response sourced from the internet). The two responses appear in a random order on each form.

The twenty academic questions and their corresponding categories are shown in the table below. These questions are designed to reflect topics likely to be discussed in a high school classroom.

Category	Question
Philosophy	Are humans good or evil?
Philosophy	Is morality relative?
Philosophy	Is it more important to be respected or liked?
Philosophy	Do humans have free will?
Philosophy	What is happiness?
Sciences	What are the states of matter?
Sciences	How do storms form?
Sciences	Which force keeps boats afloat?
Sciences	How are rainbows made?
Sciences	What enables airplanes to fly?
History	Why did the Japanese bomb Pearl Harbor?
History	What factors led to the Great Depression?
History	Who was Thomas Jefferson?
History	Which treaty ended World War I?
History	What led to the decline of the Roman Empire?
Writing/Literature	How does one craft a rhetorically convincing argument?
Writing/Literature	What is the difference between tone and mood?
Writing/Literature	How does one effectively analyze poetry?
Writing/Literature	What are the stages of a coming-of-age story?
Writing/Literature	What is an exposition in literature?

ChatGPT responses were sourced simply by pasting each question into the prompt on a new thread and then copying the AI model's response. Human-authored responses were sourced by inputting each question into the Google search engine and then copying a section of the first web page in the search results that provides attribution to a human author. The extra precaution of

requiring an author to be listed on the web page was taken to prevent using another AI generated response for being mistaken as written by a human. The text of an example form is displayed below.

Guidelines:

- Read the following question and the two responses to the question
- One of these responses was written by ChatGPT, a generative AI tool which mimics human writing styles and harnesses information from a wide array of internet sources. The other response was written by a human and published on the internet.
- Without using any outside aides (e.g. another person, Google, an AI detector, etc.) respond to the following two questions based on your reading of the responses

Question: Which treaty ended World War I?

Response 1:

Versailles: June 28, 1919

Signed in the Hall of Mirrors at the Palace of Versailles outside Paris exactly five years after the assassination of Archduke Franz Ferdinand that sparked the war, this treaty designated Germany responsible for losses and damages from the war.

Among the terms, Germany was ordered to pay crippling war damages, which became a source of deep national resentment that fuelled Nazi propaganda and is seen as a key underlying cause of the second world war.

Germany was divided by the Danzig corridor which isolated eastern Prussia from the rest of the country. Berlin lost 15 percent of its territory and 10 percent of its population.

The regions of Alsace and Lorraine were returned to France and the Saarland region in Germany was placed under an international mandate for 15 years.

The treaty was never ratified by the United States Senate.

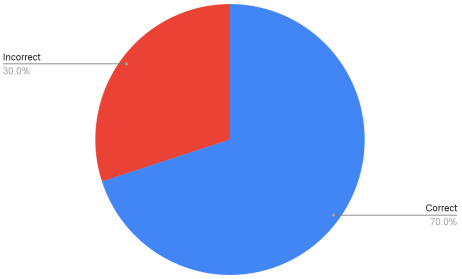
Response 2:

The treaty that ended World War I was called the Treaty of Versailles. It was signed on June 28, 1919, in Versailles, France, and officially ended the state of war between Germany and the Allied Powers (primarily composed of Britain, France, Italy, and later joined by the United States). The treaty laid down the terms and conditions for peace, including territorial adjustments, disarmament provisions, war reparations, and the establishment of the League of Nations, an international organization aimed at maintaining peace and preventing future conflicts.

In this form, Response 2 was generated by ChatGPT and Response 1 was sourced from an article on World War I in the South China Morning Post. Subsequently, each form prompts respondents to answer the following two questions:

1. Which response do you think was generated by ChatGPT?
2. Based on the overall quality of the responses, which response do you feel better answers the question?

III. Data



30 of the 90 teachers at East Chapel Hill High responded to one of the forms. Of the 30 respondents, 21 correctly identified the response generated by ChatGPT, yielding a proportion of 0.7.

The proportion of teachers who correctly identified the response generated by ChatGPT for each segment is listed below:

Philosophy: 0.67 Sciences: 0.75 History: 0.625 Writing/Literature: 0.71

Moving on to the quality of responses, 17 of the 30 teachers (57%) identified the response generated by ChatGPT as the response that better answers the question. Furthermore, 20 of the 30 teachers (67%) chose the response that they thought was human-authored as the one they felt better answered the question, alluding to possible bias towards human-authored responses that will be explored later.

IV: Statistical Methodology

Below is an outline of the necessary steps to construct a confidence interval for the proportion of teachers who can correctly identify which of two responses to an academic question was generated by ChatGPT.

STATE THE STATISTICAL OBJECTIVE: Construct a 95% confidence interval for the proportion of teachers who can correctly identify which of two responses to an academic question was generated by ChatGPT.

CHECK CONDITIONS:

Random: This study was designed to use a cluster sample of a single school rather than a random sample of teachers across the United States. This sample choice was chosen for practicality and because the teachers at one school should be representative of the entire population of teachers as long as certain factors like education, experience with ChatGPT, and subject distribution are consistent with national averages. Nonresponse bias is discussed in section VI.

Normal: $n\hat{p} = 30(0.7) = 21 \quad n(1-\hat{p}) = 30(0.3) = 9$

The expected value of teachers that do not identify the ChatGPT response is one less than 10, meaning the normal condition is not satisfied. Given the closeness to the cut-off, the inference procedure is continued with caution.

Independent: The use of twenty different forms and the withholding of correct answers from teachers ensures that individual responses are independent from each other.

CONSTRUCT THE INTERVAL

Confidence Interval for Proportions = $\hat{p} \pm z^* *$ SE

$z^* = 1.96$ for 95% confidence

SE $= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.7(0.3)}{30}} = 0.084$

Confidence Interval $= 0.7 \pm 1.96 * 0.084 = 0.7 \pm 0.17 = (0.53, 0.87)$

INTERPRET

One can be 95% confident that the true proportion of high school teachers that can correctly identify which response was generated by ChatGPT when presented with two responses to an academic question is on the interval (0.53,0.87).

The same process can be used to determine a confidence interval for the proportion of high school teachers that feel ChatGPT better answers an academic question in the survey environment previously described. This process yields a confidence interval of (0.39,0.74).

Furthermore, to inspect whether a teacher's selection of the response they feel better answers the given question is independent of the choice they perceive to be generated by ChatGPT, Fisher's Exact Test* is executed below.

Contingency Table:

	Chose the Response Generated by ChatGPT as the one that better answers the question	Chose the human-authored response as the one that better answers the question	TOTAL
Correctly Identified the Response Generated by ChatGPT	a = 9	b = 12	21
Incorrectly Identified the Response Generated by ChatGPT	c = 8	d = 1	9
TOTAL	17	13	30

For this procedure, the default significance level of $\alpha = 0.05$ is used. Fisher's Exact Test necessitates the independence of observations which is established earlier in this section. The P-value can is computed using the following formula:

$$\text{P-value} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{21!9!17!13!}{9!12!8!1!30!} = 0.022$$

Since 0.022 is less than α, the null hypothesis that the teachers' perception of quality is independent of which response was perceived to be generated by ChatGPT is rejected. At the 0.05 significance level, one can accept that the response a teacher feels better answers a given question is dependent on which one they perceive to be written by ChatGPT.

V: Statistical Analysis

The confidence level for the proportion of teachers that can recognize the passage written by ChatGPT ranges from 0.53 to 0.87, meaning that the entire range is above 0.5. Since 0.5 represents the resulting proportion if teachers have no intuition as to which response was generated by ChatGPT, one can conclude that teachers have at least some ability to detect ChatGPT Responses. The margin of error is fairly large on the sample proportion obtained in this study (largely as a result of the relatively small sample size), causing the ability of the population of high school teachers to detect ChatGPT to be largely indeterminate.

Looking at the sample proportions of ChatGPT responses recognized divided on category (Philosophy: 0.67, Sciences: 0.75, History: 0.625, Writing/Literature: 0.71), all of these proportions are relatively close to the sample proportion for the entire sample (within 0.075). New data with larger sample sizes would need to be collected to perform z-tests for difference between proportions, and detect if there is a statistically significant difference in recognition due to subject area.

Furthermore, the confidence interval for the proportion of teachers that feel that the ChatGPT response better answers the question ranges from 0.39 to 0.74. Since 0.5 is included in this interval, one cannot conclude whether the population of high school teachers feels that ChatGPT responses or human-authored responses better answer academic questions.

Lastly, the Fisher's exact test shows that which response a teacher feels better answers a question is dependent on which response they feel was written by ChatGPT. This dependence could be the result of several factors, namely preconceived notions on the quality of ChatGPT writing and bias against selecting ChatGPT as higher quality. The fact that only one of the nine teachers that did not correctly identify the response generated by ChatGPT selected the response they perceived to be written by ChatGPT as the one that is higher quality suggests that teachers who fail to recognize the use of ChatGPT may do so because they think that the AI model has lower quality of writing than it actually does.

VI: Discussion on Error and Inconsistencies with a Classroom Environment

Ideally, a large random sample of high school teachers across the United States would be gathered to execute this study. However, generating such a sample is not at all practical. Instead, a cluster sample of high schools is a valid alternative. This study uses a single cluster, which is not an ideal sample since there is definite possibility that the chosen cluster is not representative of the population at large. In the case of East Chapel Hill High, above average teacher education and experience with ChatGPT may contribute to an increased ability to detect its use.

Another flaw in this study is the nonresponse bias. Only 30 of the 90 teachers asked to participate in this study responded. Among other factors, the teachers that chose not to participate may have chosen not to participate because they feel less comfortable identifying the use of ChatGPT or less interested in the matter altogether. If these proposals are reasons behind the low response rate, the confidence interval may need to be adjusted lower.

During the study, teachers voiced concern over whether the survey is reflective of a classroom environment. In a classroom setting, teachers are familiar with the quality of writing and writing style of their students, making it easier to detect a deviation from past written works. Furthermore, teachers are aware of what content has and has not been covered in their class. If a student submits a piece of work that covers topics not discussed in class, this can indicate that a student used a generative AI tool to complete their assignment. Both of these differences between the survey and classroom suggest that the confidence interval for ChatGPT recognition may need to be adjusted upwards.

VII: Conclusion

This study uses confidence intervals and Fisher's exact test to analyze the ability of high school teachers to detect the use of ChatGPT and high school teachers' perception of the quality of ChatGPT responses. Design flaws and a high margin of error on confidence intervals as a result of a small sample size prohibit any precise conclusions. That being said, this study was successful in concluding that teachers at least have some ability to detect ChatGPT's use. Furthermore, the Fisher's Exact test demonstrated that a teacher's perception of which response better answers the question depends on which response they feel was generated by ChatGPT. A larger randomized trial at a national level is suggested to solidify some of the ideas suggested in this paper.

VII: References (including sources used for forms)

OpenAI. "ChatGPT: AI Language Model." OpenAI, 2021.

Gordon, Cindy. “How Are Educators Reacting to Chat GPT?” *Forbes*, 2 May 2023, www.forbes.com/sites/cindygordon/2023/04/30/how-are-educators-reacting-to-chat-gpt/?sh=12f4155e2f1c.

Leigh, Dana. “How Does Chat GPT Actually Work?” *TechRound*, 11 Apr. 2023, techround.co.uk/guides/how-does-chat-gpt-actually-work/.

Zach. “Fisher’s Exact Test: Definition, Formula, and Example.” *Statology*, 9 July 2020, www.statology.org/fishers-exact-test/.

Aglietti, Tom. “Are We Born Good or Evil?” BBC Earth, 1996, www.bbcearth.com/news/are-we-born-good-or-evil-naughty-or-nice.

Manuel Velasquez, Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer. “Ethical Relativism.” Markkula Center for Applied Ethics, 1 Aug. 1992, www.scu.edu/ethics/ethics-resources/ethical-decision-making/ethical-relativism/#:~:text=Ethical%20relativism%20is%20the%20theory,be%20morally%20wrong%20in%20another.

Klessinger, John. “Being Liked vs. Being Respected.” Coaches Insider, 8 Sept. 2021, coachesinsider.com/soccer/being-liked-vs-being-respected-10/.

O’Connor, Timothy, and Christopher Franklin. “Free Will.” Stanford Encyclopedia of Philosophy, 3 Nov. 2022, plato.stanford.edu/entries/freewill/.

Cherry, Kendra. “How Do Psychologists Define Happiness?” Verywell Mind, 7 Nov. 2022, www.verywellmind.com/what-is-happiness-4869755.

Bagley, Mary. “States of Matter: Definition and Phases of Change.” LiveScience, 20 Oct. 2022, www.livescience.com/46506-states-of-matter.html.

Franklin-Cheung, Alexandra. “How Do Thunderstorms Form?” BBC Science Focus Magazine, 24 Oct. 2020, www.sciencefocus.com/planet-earth/how-do-thunderstorms-form/.

Farid, Hany. “Why Do Boats Float and Rocks Sink? .” ReadWorks, 2015, www.psd1.org/cms/lib/WA01001055/Centricity/Domain/1665/Density.pdf.

Mahlen, Gena. “How a Rainbow Is Formed.” How Are Rainbows Made?, faculty.cord.edu/manning/physics215/studentpages/genamahlen.html. Accessed 18 May 2023.

Benson, Tom. “What Makes a Plane Go Up?” NASA, 13 May 2021, www.grc.nasa.gov/www/k-12/Summer_Training/Elementary97/planearticle.html.

Alghussein, Jason. “Why Japan Attacked Pearl Harbor.” Pearl Harbor, 11 Feb. 2023, pearlharbor.org/blog/why-japan-attacked-pearl-harbor/.
Duignan, Brian. “Causes of the Great Depression.” Encyclopædia Britannica, 2023, www.britannica.com/story/causes-of-the-great-depression.

Ellis, Joeseph. “Thomas Jefferson.” Encyclopædia Britannica, 9 Apr. 2023, www.britannica.com/biography/Thomas-Jefferson.

France-Presse, Agence. “WWI Centenary: 7 Peace Treaties That Ended the First World War, from Versailles to Lausanne.” Young Post, 10 Nov. 2018, www.scmp.com/yp/discover/lifestyle/features/article/3058235/wwi-centenary-7-peace-treaties-ended-first-world-war.

Andrews, Evan. “8 Reasons Why Rome Fell.” History.Com, 14 Jan. 2014, www.history.com/news/8-reasons-why-rome-fell.

Lloyd, Amanda, et al. “6.4 Rhetorical Appeals: Logos, Pathos, and Ethos Defined.” A Guide to Rhetoric Genre and Success in FirstYear Writing, 2023, pressbooks.ulib.csuohio.edu/csu-fyw-rhetoric/chapter/rhetorical-strategies-building-compelling-arguments/.

Stone, Lucia, and Marco Norris . “What Is the Difference between Mood and Tone? || Definitions and Examples.” College of Liberal Arts, 6 Feb. 2023, liberalarts.oregonstate.edu/wlf/what-difference-between-mood-and-tone-definitions-and-examples#:~:text=Mood%20shows%20the%20particular%20scenes,actually%20thinks%20of%20that%20subject.

Miller-Wilson, Kate. “How to Analyze a Poem Effectively.” YourDictionary, 21 July 2020, grammar.yourdictionary.com/writing/how-to-analyze-a-poem-effectively.html.

Littlehale, Kristy. “Bildungsroman Definition & Examples: Coming of Age Novels.” Storyboard That, 2023, www.storyboardthat.com/articles/e/bildungsroman-novels.

Fleming, Grace. “What Is Exposition in Literature? Check out These Popular Examples.” ThoughtCo, 10 Aug. 2019, www.thoughtco.com/what-is-exposition-1857641.

Appendix

*Traditionally, a chi-square test for independence would be used in this situation. However, the condition on expected values is not met, thus making Fisher’s Exact Test the correct procedure. Fisher’s exact test calculates the exact probability of a table arrangement as extreme or more extreme than a given contingency table with extremity being a measure of how far the table deviates from being evenly distributed.