[Result, but poetic]: Does LLM Use Improve Data Science Education?*

Evidence from a Canadian Undergraduate Statistics Course

Rohan Alexander   Luca Carnegie

July 30, 2024

An abstract should eventually go here!!

1 Introduction

The wide release of ChatGPT and other Large Language Models (LLMs) has rapidly transformed various parts of society, in particular the field of education. As these tools are increasingly accessible, their adoption and use in academic settings has created excitement and apprehension among educators and students alike. LLMs offer unprecedented support for students in tasks ranging from writing assistance to support in complex problem-solving, with the potential to enhance academic performance and student outcomes. At the same time, their use raises questions about academic integrity, the development of critical thinking skills (particular in undergraduates), as well as questions about their implications for effective learning overall.

The potential for LLMs to positively affect students' academic performance in data science can be clearly inferred from already-demonstrated effects on adjacent professional fields. The tasks involved in a data science workflow can be generally broken down into two key competencies. The first is programming, which is done when cleaning, analyzing, visualizing data using languages like R or Python. The second is writing, which is primarily done when communicating results. There is experimental evidence from adjacent professional fields that LLMs can improve productivity in both of these competencies.

Peng et al. (2023) show the potential impact of LLMs on students' computer programming skills, through their uncovering of the significantly positive impacts of GitHub Copilot (an LLM-powered programming assistant) on the productivity of computer programmers. In an experiment involving 95 freelance programmers, they found that that programmers with access

*Code and data are available at: https://github.com/lcarnegie/llms-achievement.

to Copilot completed a standardized programming task 55.8% faster than the control group. Crucially, it was found that programmers with less experience saw the greatest improvements in productivity. Given that programming is a core component of data science practice, this suggests that these benefits could reasonably carry over to students in data science courses as well.

Dell'Acqua et al. (2023), while primarily about management consulting tasks, provides evidence that LLMs can significantly improve writing productivity. In a field experiment involving Consultants recruited from the Boston Consulting Group (BCG), they found that the use of GPT-4 in the experimental group led to a 25% increase in delivery speed of business tasks (most involving some writing), as well as a 40% increase in human-rated performance on those tasks. Similar to computer programming, these productivity increases were most pronounced for those with below average performance, with their output increasing by 43%. Though not all tasks done using AI saw the same productivity improvements, the authors expressed particular optimism for LLMs' potential to generally expedite menial knowledge-work tasks. These included tasks such as generating new ideas and creating persuasive and informative writing - both examples directly apply to data science education through their analogues of coming up with project ideas and writing more engaging reports. As a whole, they show that LLMs could equally augment the quality of student writing and course projects in data science as well.

Clearly, the implications of these case studies appear promising. However, literature from educators both adjacent to and distant from statistics and data science suggest much more varied perspectives on how LLMs can impact student performance.

Valenzuela et al. (2024) argue that LLMs lead to a loss of serendipity (which leads to less original work) and de-skilling (primarily with respect to programming ability), among other consequences. These particular outcomes could negatively affect students' effective learning of data science. On the other hand, Ellis and Slade (2023) take a more optimistic perspective, taking the popular stance of comparing ChatGPT to previously controversial learning technologies like calculators. They argue that LLMs are just another technology that will impact Statistics and Data Science education like the calculator did. However, they take a broader perspective on the issue and more specific inquiry could be done in how effective LLMs are at improving student outcomes within data science education in particular.

In a similar attitude, Tu et al. (2024) acknowledge that for students, LLMs can streamline many parts of a data science workflow - with that in mind, they suggest that data scientists in training should shift their self-perspectives from primarily being an analyst to primarily being a manager responsible for strategic oversight of the analysis. Crucially they emphasize that in both education and practice, LLMs and human intelligence should play complementary roles.

The simultaneous caution and interest the literature expressed toward using LLMs in educational settings was further corroborated by empirical studies on K-12 and university students as well.

At the high school level, Lazar et al. (2023) conducted a informal survey of secondary school teachers and students on their opinions of ChatGPT, they found that while LLMs could help spark creativity, provide academic support when teachers were unavailable, and model certain types of writing well, teachers were also cognizant of LLMs potential to limit students' learning in certain ways through overreliance. Beyond academic integrity concerns, teachers had similar concerns to Valenzuela et al. (2024) about de-skilling and an overall loss of agency in writing and critical thinking.

Cahill and McCabe (2024) surveyed undergraduate Political Science students on their attitudes toward and usage of AI tools. They found that the use of ChatGPT (among other machine-learning-powered software) was widespread. However, they also found that many students lacked the confidence in using AI for academic purposes - in particular, only 11% 'strongly agreed' that they know how to use AI to improve their writing. Like educators, students have nuanced views on appropriate AI use. In particular, respondents found that using it to writing whole papers as inappropriate, while using it for basic tasks like general assistance, writing feedback and basic data visualization was perceived more appropriately. Interestingly, first-generation college students were found to be more likely to use AI in their work, particularly in writing papers and helping with assignments. Their findings suggest that LLMs and other AI tools could be an equalizer for disadvantaged students. Rhough only political science students were surveyed, statistics and data science students could reasonably have general attitudes that are similar.

As we have seen, the existing inquiry by educators has explored the general qualitative usage and student perceptions on these tools. Though informative, there is still a key lack of evidence of how students' academic performance can be affected through allowing them to use LLMs in classwork. This paper aims to fill this gap by quantitatively investigating the relationship between students' grades and measures of student LLM usage and their attitudes toward LLMs in general, using evidence from a third-year undergraduate statistics course at the University of Toronto. Through examining how students interact with and perceive LLM tools and how these variables translate into student outcomes, the effects of LLM integration in data science education can be more precisely determined.

The remainder of this paper is structured as follows: Section 2 visualizes and analyzes survey data and coursework from students; Section 3 models the unstructured data to approximate a relationship between grades and usage/attitudes; Section 4 describes and analyzes the model's results; Section 5 discusses the implications of the findings for data science education and future research and practice in this rapidly evolving field.

Table 1: (CHANGE) Boilerplate table caption

2 Data

2.1 Background and Collection

To investigate students' usage and attitudes toward LLM use and how they related to their academic performance, a dataset containing their usage/attitudes, coursework, and academic performance was constructed. Data was sampled from the cohort of students taking STA302 - Methods of Data Analysis I, taught by one of the investigators in the Winter 2024 semester at the University of Toronto. By virtue of pre-requisites needed, this restricted the data collected to be only on upper-year undergraduate students.

Data was collected from students through an optional end-of-course survey. Whether or not they consented to their data being used in this investigation, all respondents received a +1% increase in their final course grade for their participation. All consenting responses were then cross-referenced to their course grade, as well as the GitHub account they used to complete their course research papers. The responses were finally anonymized by removing any personal references to the students themselves.

2.2 Cleaning and Variables of Interest

All the data was cleaned using R (R Core Team 2023) and it's tidyverse (Wickham et al. 2019), janitor (Firke 2023) and stringr (Wickham 2023), then tested for issues using the testthat package (Wickham 2011). This led to a final dataset containing 121 responses across (var_num) variables.

These variables were:

- [list of variables]

2.3 Visualization

The data was then visualized using the ggplot2 (Wickham 2016) package.

Maybe a table for the years instead?

Get inspired by Cahill and McCabe's paper for visualizations etc.

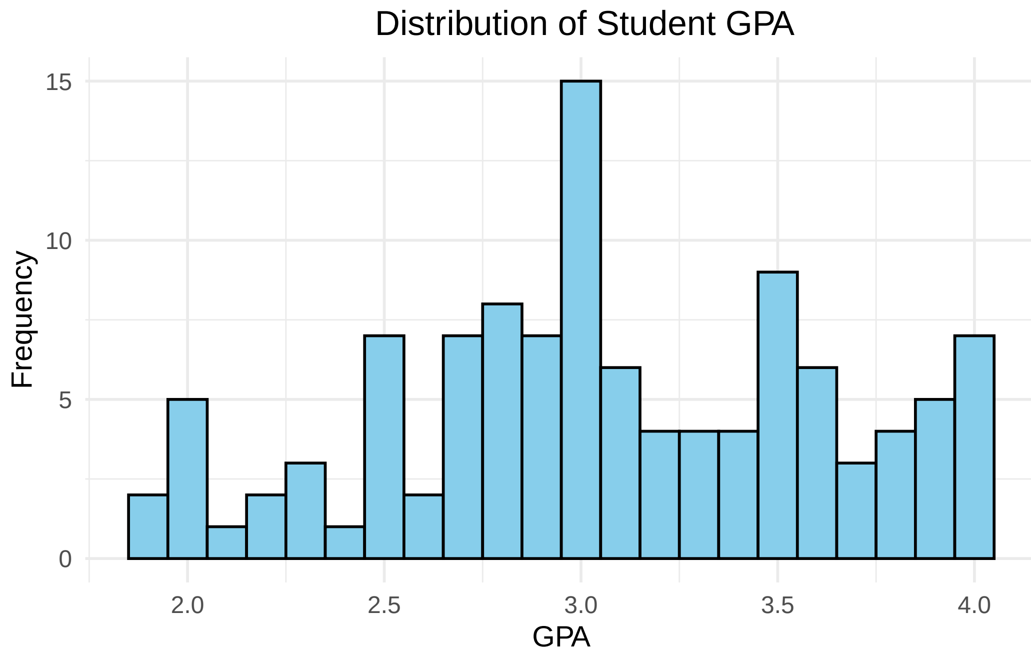

Figure 1

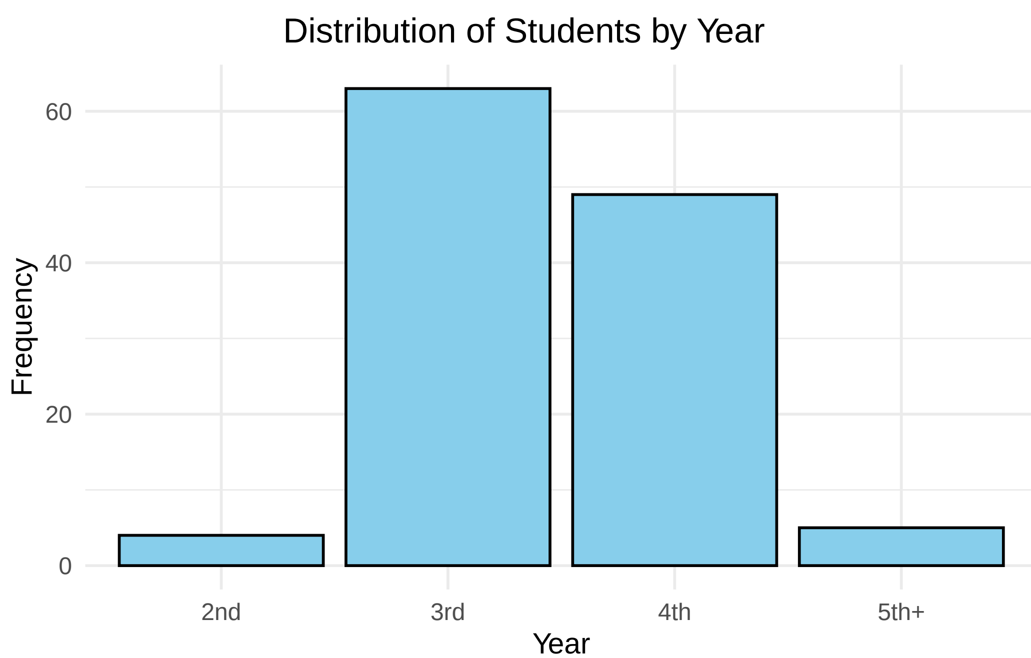

Figure 2

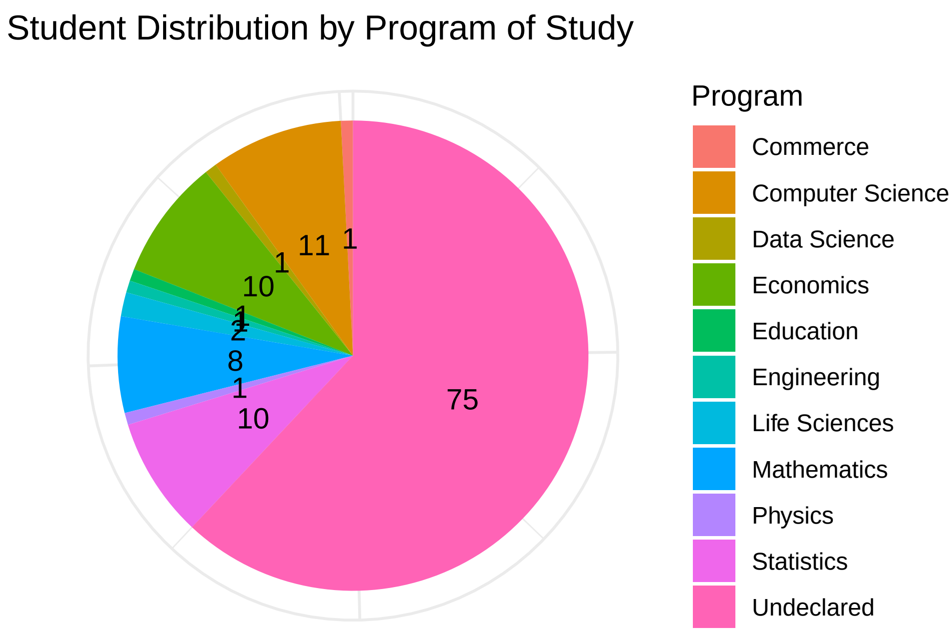

Figure 3

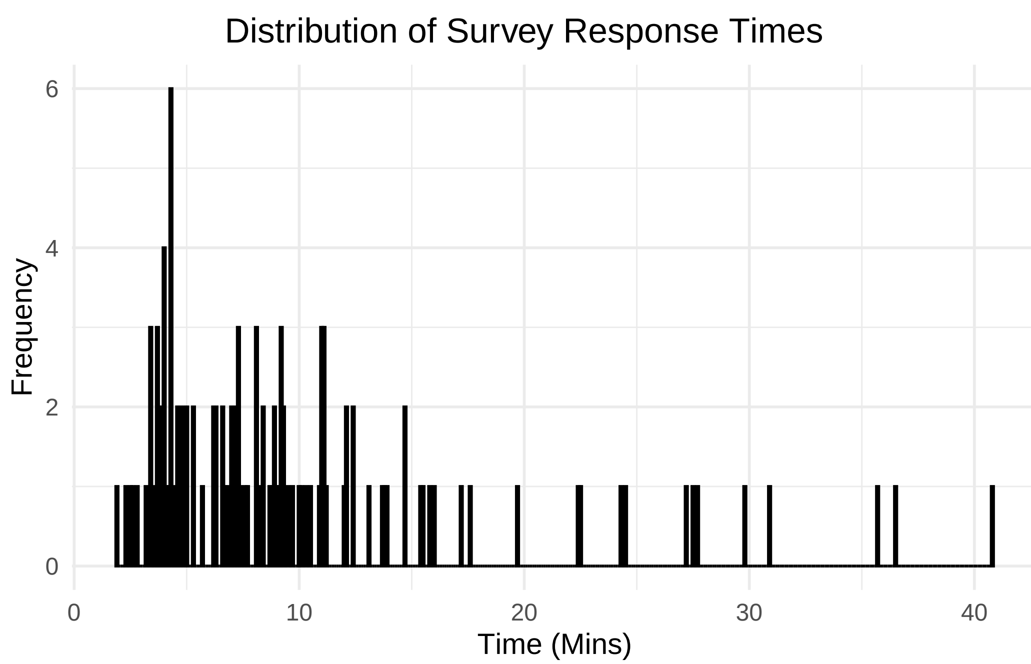

Figure 4: Distribution of Survey Response Times

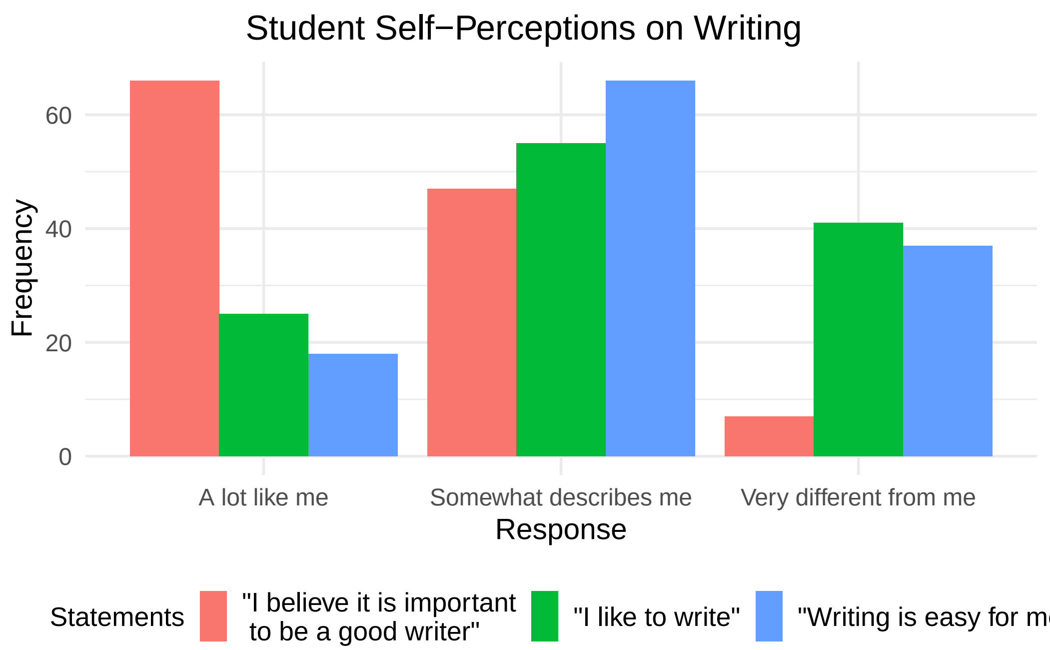

Figure 5

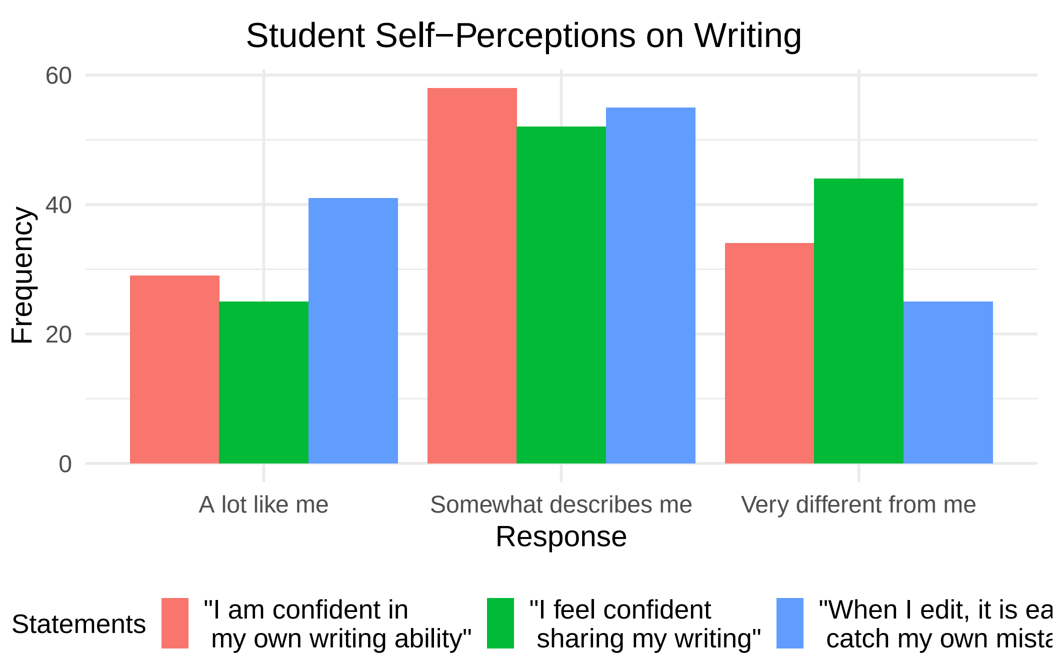

Figure 6

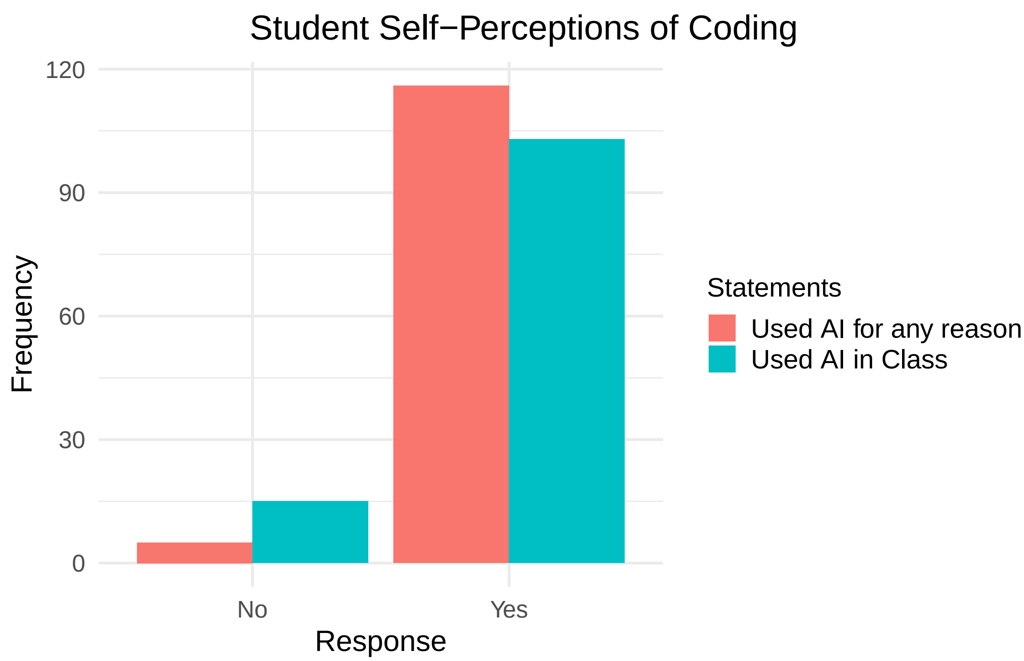

Figure 7: Hello

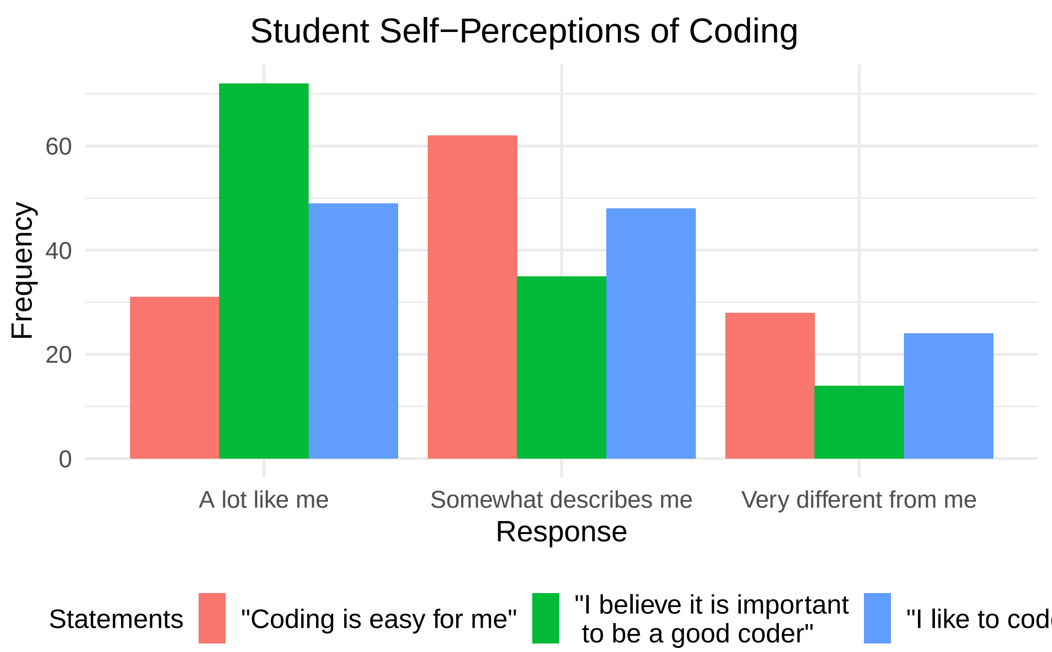

Figure 8

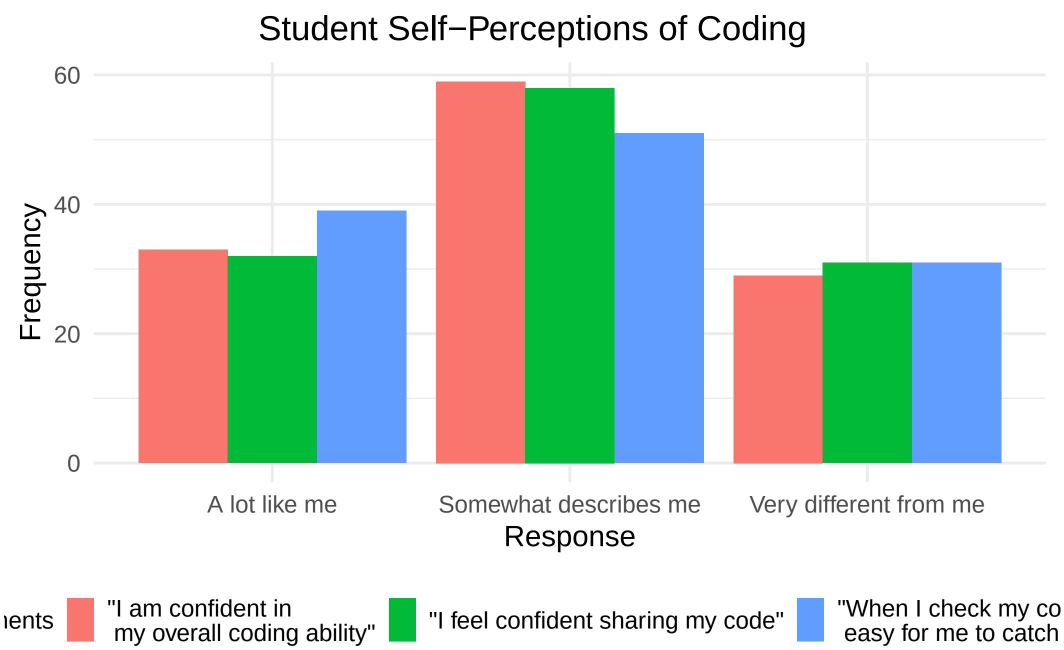

Figure 9

Table 2: Familiarity with AI of students

Not familiar	Somewhat familiar	Very familiar
3	65	53

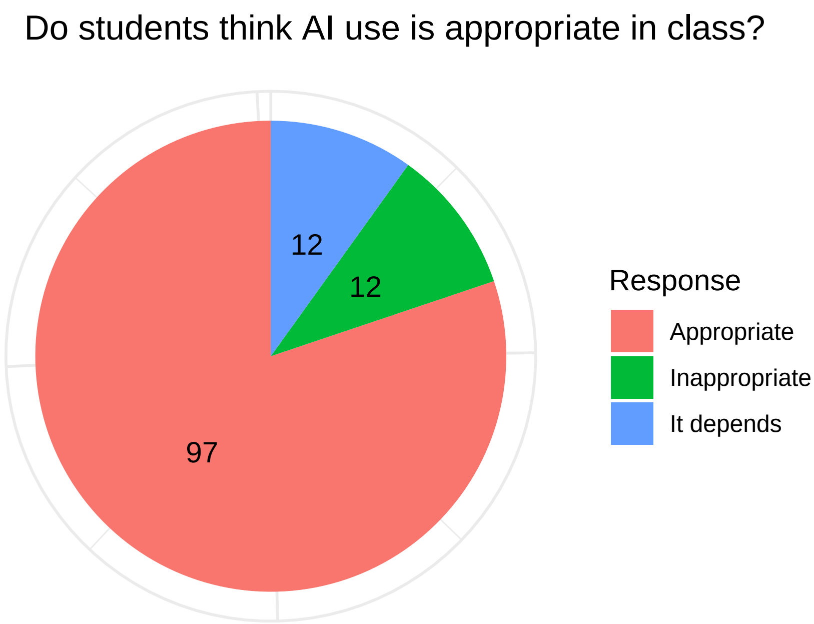

Figure 10

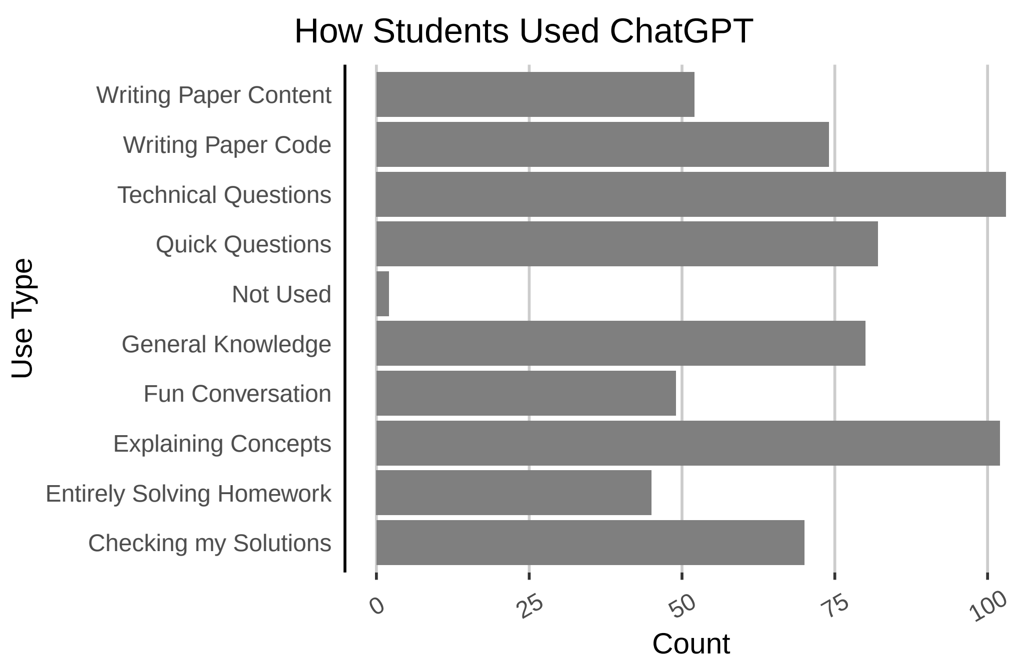

Figure 11: How Students Use ChatGPT

2.4 Sentiment Analysis

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In ... we implement a posterior predictive check. This shows...

In ... we compare the posterior with the prior. This shows...

B.2 Diagnostics

... is a trace plot. It shows... This suggests...

... is a Rhat plot. It shows... This suggests...

References

Ben-Michael, Eli, D James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin. 2024. "Does AI Help Humans Make Better Decisions? A Methodological Framework for Experimental Evaluation." *arXiv Preprint arXiv:2403.12108*.

Cahill, Christine, and Katherine McCabe. 2024. "Context Matters: Understanding Student Usage, Skills, and Attitudes Toward AI to Inform Classroom Policies." *PS: Political Science & Politics*, May, 1–8. https://doi.org/10.1017/S1049096524000155.

Carobene, Anna, Andrea Padoan, Federico Cabitza, Giuseppe Banfi, and Mario Plebani. 2024. "Rising Adoption of Artificial Intelligence in Scientific Publishing: Evaluating the Role, Risks, and Ethical Implications in Paper Drafting and Review Process." *Clinical Chemistry and Laboratory Medicine (CCLM)* 62 (5): 835–43. https://doi.org/10.1515/cclm-2023-1136.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4573321.

Ellis, Amanda R., and Emily Slade. 2023. "A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education." *Journal of Statistics and Data Science Education* 31 (2): 128–33. https://doi.org/10.1080/26939169.2023.2223609.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Kross, Sean, and Philip J. Guo. 2019. "Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://api.semanticscholar.org/CorpusID:102493683.

Lazar, Nicole, James Byrns, Danielle Crowe, Meghan McGinty, Angela Abraham, Mike Guo, Megan Mann, et al. 2023. "Perils and Opportunities of ChatGPT: A High School Perspective." *Harvard Data Science Review* 5 (4).

Li, You, Ye Wang, Yugyung Lee, Huan Chen, Alexis Nicolle Petri, and Teryn Cha. 2023. "Teaching Data Science Through Storytelling: Improving Undergraduate Data Literacy." *Thinking Skills and Creativity* 48 (June): 101311. https://doi.org/10.1016/j.tsc.2023.101311.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," no. arXiv:2302.06590 (February). http://arxiv.org/abs/2302.06590.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Truncano, Michael. 2023. "AI and the Next Digital Divide in Education | Brookings." *Brookings*. https://www.brookings.edu/articles/ai-and-the-next-digital-divide-in-education/.

Tu, Xinming, James Zou, Weijie Su, and Linjun Zhang. 2024. "What Should Data Science Education Do With Large Language Models?" *Harvard Data Science Review* 6 (1).

Valenzuela, Ana, Stefano Puntoni, Donna Hoffman, Noah Castelo, Julian De Freitas, Berkeley Dietvorst, Christian Hildebrand, et al. 2024. "How Artificial Intelligence Constrains the Human Experience." *Journal of the Association for Consumer Research* 9 (3): 241–56. https://doi.org/10.1086/730709.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.