Self-reported LLM usage and results on a data science project: Evidence from a Canadian undergraduate data science course*

Rohan Alexander Luca Carnegie Nathalie Moon

August 14, 2024

To help understand the effect of Large Language Models (LLMs) on data science practice we examine the extent to which LLM usage is correlated with the mark that a student gets on a final paper in a classroom data science setting. We find some very mild evidence from this observational study that LLM usage may be associated with better scores, but our main conclusion is that there is no clear relationship between more extensive LLM usage and the student's mark. Despite the classroom setting used for evaluation, the particular activity of interest is similar to the work done by professional data scientists. Our finding suggests the need for more extensive work evaluating how LLMs can be integrated into the data science workflow in a way that provides value.

1 Introduction

Trustworthy data science is the practice of conducting data analysis in a transparent, ethical, and reliable manner. These principles are upheld in various ways, through ongoing education, adherence to professional norms and, and implementation of reproducible workflows. To stay current, data scientists must continually update their knowledge of best practices for transparency and reproducibility, foster a culture which values different perspectives and

*Code and data are available at: https://github.com/lcarnegie/llms-achievement. We thank Tiffany Timbers and attendees at JSM 2024 for helpful suggestions. This research is currently under review by the University of Toronto's Research Ethics Board. Contributions: RA: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; LC: Formal Analysis, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; NM: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing. Comments can be sent to: rohan.alexander@utoronto.ca.

accountability, and critically consider the broader impact of data-driven decisions. Recent advances, such as the wide release of user-friendly Large Language Models (LLMs), particularly OpenAI's ChatGPT, have transformed the data science toolkit. These powerful tools for natural language processing and generation influence various aspects of data analysis and automation, further emphasizing the need for trustworthy practices.

Like all new tools LLMs have created both excitement and apprehension. ChatGPT's public release on November 30, 2022, brought LLMs into the mainstream conversation. LLMs have quickly been put to use in both industry and academia. By now, many people, especially educators and students, have some experience with LLMs, in both personal and professional contexts.

In the context of teaching statistics and data science, LLMs could be useful for many tasks. For instance, there is considerable interest in the potential of chatbots to act as personalized tutors for students (Fulgencio 2024; Afzal, Zamir, and Ali 2024). ChatGPT has also been the catalyst for many interesting and important conversations around academic integrity (Eke 2023), the development of critical thinking skills (Thiga 2024), and what effective learning looks like (Baidoo-Anu and Ansah 2023).

In this paper we are interested in better understanding LLMs as a tool for producing trustworthy data science. We study how they were used by students in an upper-year undergraduate data science course and whether students who used LLMs tended to have higher scores than those who did not.

The tasks involved in a trustworthy data science workflow can be generally broken down into a number of key competencies (Adhikari, DeNero, and Jordan 2021; Gibbs and Taback 2021). One is programming, which is done when cleaning, analyzing, and visualizing data often using programming languages like R or Python. Another is writing, which is primarily done when communicating results. The potential for LLMs to positively affect students' academic performance in data science can be clearly inferred from their already-demonstrated capability in those competencies in adjacent professional fields.

In terms of computer programming, Peng et al. (2023) found a positive impact of GitHub Copilot (an LLM-powered programming assistant) on productivity. Specifically, in an experiment involving 95 freelance programmers, they found that a treatment group of programmers with access to GitHub Copilot completed a standardized programming task 56% faster than the control group. Programmers with less experience saw the greatest improvements in productivity.

Dell'Acqua et al. (2023), focusing on management consulting tasks, provide evidence that LLMs can improve writing productivity. In a field experiment involving consultants from Boston Consulting Group they found that the use of OpenAI's GPT-4 led to a 25% increase in delivery speed of business tasks, most of which involved 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%.

On the other hand, Valenzuela et al. (2024) argue that LLMs lead to a loss of serendipity which may lead to less original work, and potentially de-skilling primarily with respect to programming ability, among other consequences. These outcomes could negatively affect students' effective learning of data science.

Ellis and Slade (2023) take a more optimistic perspective and argue that LLMs are just another technology that will impact statistics and data science education. Similarly, Tu et al. (2024) acknowledge that 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 "product manager" responsible for strategic oversight of the analysis carried out by LLMs.

At the high school level Lazar et al. (2023) conducted an informal survey of secondary school teachers and students on their opinions of ChatGPT, and found that while LLMs could: help creativity, provide academic support when teachers were unavailable, and model certain types of writing well. Teachers were also concerned about the potential of LLMs to limit students' learning in certain ways through over-reliance. Beyond academic integrity concerns, Lazar et al. (2023) found that 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 was widespread. However, they also found that many students lacked confidence in using AI for academic purposes. In particular, only 11% "strongly agreed" that they know how to use AI to improve their writing. Students had nuanced views on appropriate AI use. Many respondents felt that using it to write whole papers was inappropriate, but using it for basic tasks like general assistance, writing feedback and basic data visualization appropriate.

To understand the current state of LLMs as a tool for trustworthy data science, this paper focuses on the association between student academic performance and their LLM usage. Specifically, we examine the relationship between students' grades and self-reported measures of student LLM usage, as well as student attitudes toward LLMs in general. This is based on students' final papers and a survey, conducted in a third-year undergraduate data science course at the University of Toronto. By examining how students interact with and perceive LLMs as tools, and how these variables translate into student outcomes, current practice with regard to LLM integration in data science can be better understood, leading to better recommendations for their future development.

The remainder of this paper is structured as follows: Section 2 visualizes and analyzes survey data and coursework from students. Section 3 specifies a model used to investigate the relationship. 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 at the intersection of LLMs and trustworthy data science.

2 Data

2.1 Background

To investigate students' usage and attitudes towards LLMs and how they related to their academic performance, a dataset containing their usage and attitudes, coursework, and academic performance was constructed. This was based on three components:

1. an optional survey;
2. self-reported LLM usage; and
3. student marks on their final paper.

All data are from the cohort of students taking STA302 "Methods of Data Analysis I" in the Winter 2024 semester at the University of Toronto. This course had 275 students initially enrolled which, reflecting a normal rate of attrition for undergraduate statistics courses at the University of Toronto, reduced to 154 students by the end of the semester. Assessment was heavily based on three papers submitted over the course of the 12 week semester.

The student marks that we analyze are based only on the final paper, which is done individually. By this stage, uninterested students have typically dropped the course, and students are familiar with course expectations. A typical paper submission is 10-20 pages, and requires students to conduct original research to answer a research question of interest to them. It reflects the skills typically used by a professional data scientist. Students are expected to develop a research question of interest to them, identify or collect data to answer the question, conduct statistical analysis, and write a short paper. Examples of final papers (shared with consent) include: Yu (2024); Su (2024); and Rochwerg (2024).

By the time they are working on their final paper, students have submitted and received feedback on two previous papers with similar requirements and rubrics to that of the final paper. Each paper has the same basic structure and expectations. Before the final paper is due, students have received feedback on all their previous work in the class (including their past papers) and there is an optional two-day period of peer review.

The pre-requisites of this course mean that the typical student is an upper-year undergraduate. Coding and writing are major parts of the course. Students are welcome to use R or Python, but the majority code in R because that is the programming language currently mostly taught in pre-requisite courses. All writing must be in English. The primary motivation for having students write three papers as the main assessment for the course is to give them the opportunity to create a public portfolio of work they can use to apply for jobs.

Throughout the semester students were encouraged to use LLMs. Formal instruction was provided twice during the semester. The first was a masterclass taught by a computer science faculty member on the ethics of using LLMs (see Horton et al. (2024) for details). The second was a masterclass taught by a TA on writing with LLMs.

Data was collected from students through an optional end-of-course survey. Appendix A details the questions asked in the survey. Whether or not they consented to their data being used, all respondents received a 1% increase in their final course grade for their participation. Consenting responses were then matched to their final paper mark, as well as the GitHub repository for their final paper. The responses were anonymized by removing any personal references to the students themselves including, names, emails, student numbers, and GitHub links.

Data cleaning and analysis was done using the `R` statistical programming language (R Core Team 2023), and the `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `reshape2` (Wickham 2007), and `readxl` (Wickham and Bryan 2023) packages.

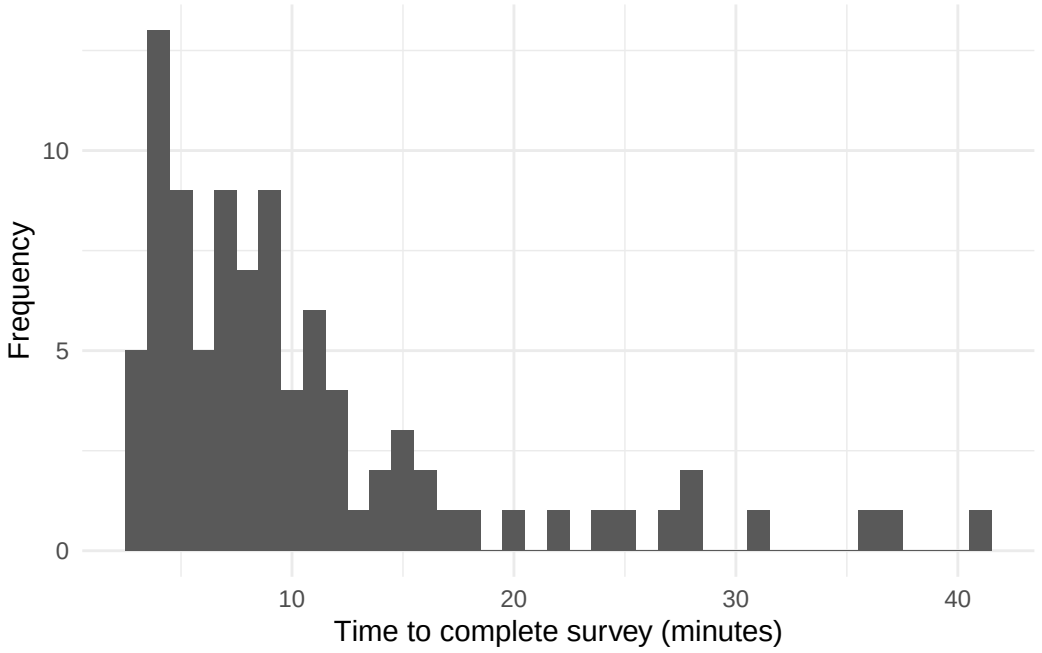
2.2 Survey data

There were 146 responses to the survey. Of these, 119 respondents provided authorization for their data to be collected and used. Four of those respondents submitted the survey twice, and after removing their second response, 115 responses remained. Of those, 15 respondents did not include a statement on LLM usage in the README of the GitHub repository of their final paper, leaving 100 responses. Finally 7 of those respondents did not provide a usable GPA response. This leaves 93 respondents that were of use and were merged based on student name.

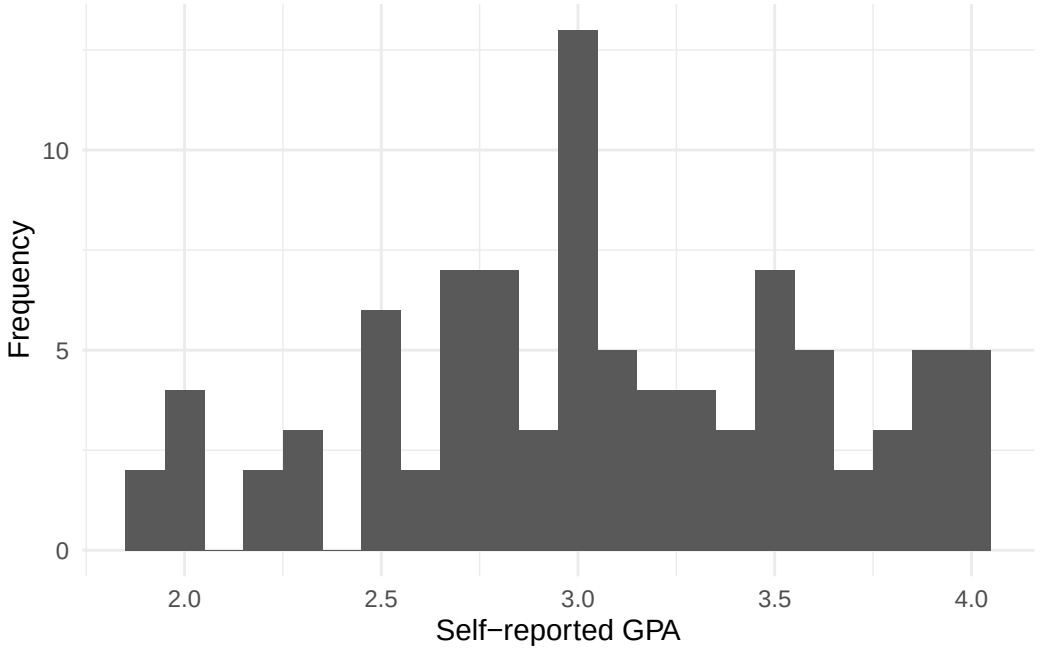
All but one respondent completed the survey within 45 minutes (Figure 1a). That one respondent took more than 4,000 minutes to complete the survey, suggesting they took a break while filling it out. That respondent is included in the analysis dataset, but without them, the average time to complete the survey was 11 minutes, and the standard deviation was 8 minutes.

There is a wide distribution of self-reported GPAs (Figure 1b). The majority of responses cluster around a B (3.0 out of 4.0), and the average is 3.06 with a standard deviation of 0.55. One factor that may affect the range is that the course is required for programs in the Statistics, Mathematics and Computer Science Departments. Self-reported GPAs also introduce the possibility of reporting bias. For instance, respondents may have provided their cumulative GPA, their most recent term's GPA, or could have misreported it entirely.

Students from a range of years took the course, however the majority of respondents were in their 3rd or 4th year of study (Table 1). The course's prerequisite two-course sequence is typically completed by students in their second year, would make it difficult to take this course earlier than their 3rd year.



(a) Distribution of survey response times



(b) "What is your GPA?"

Figure 1: Distribution of respondents' survey response times and self-reported GPA

Table 1: "What year are you?"

2nd	3rd	4th	5th or over
4	47	39	2

Respondents had a varied self-perception of their coding and writing abilities (Figure 2). Most respondents believe it is important to be good at writing, but many are either indifferent or do not like to write (Figure 2a). Respondents also do not find writing to be particularly easy, which could be associated with the reported relative antipathy toward writing. Most respondents were at least somewhat confident in their own writing abilities, but a substantial number felt otherwise. Although few respondents felt that they were confident in their writing ability, more felt that they were able to catch their mistakes, which could indicate a disconnect between how respondents perceive their work and how the work was evaluated.

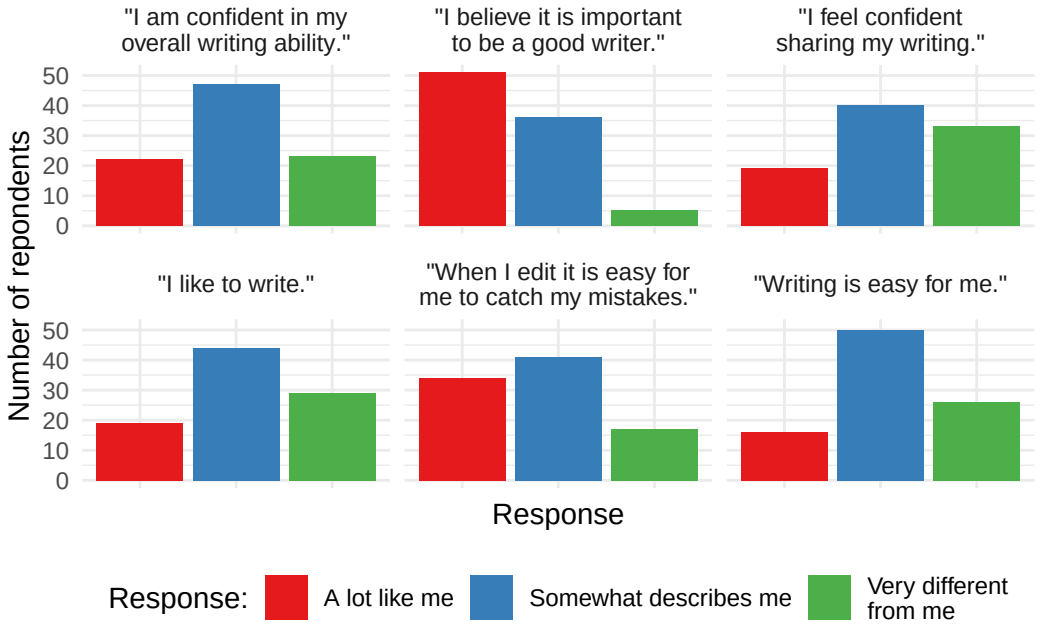
Respondent self-perceptions regarding coding proficiency and importance varied (Figure 2b). There is strong consensus on the perceived importance of coding skills. However, respondents' self-assessed coding abilities and enjoyment are more heterogeneous, with a substantial proportion reporting moderate rather than high levels of ease and enjoyment in coding tasks.

Overall there was a moderate level of confidence among respondents in their overall coding ability, willingness to share code, and capacity to identify errors (Figure 2b). Notably, respondents express slightly higher confidence in detecting their own coding mistakes compared to general coding ability or code sharing. These patterns suggest that while respondents have developed some coding self-efficacy, there is still considerable potential for enhancing their perceived competence and comfort across various coding-related activities.

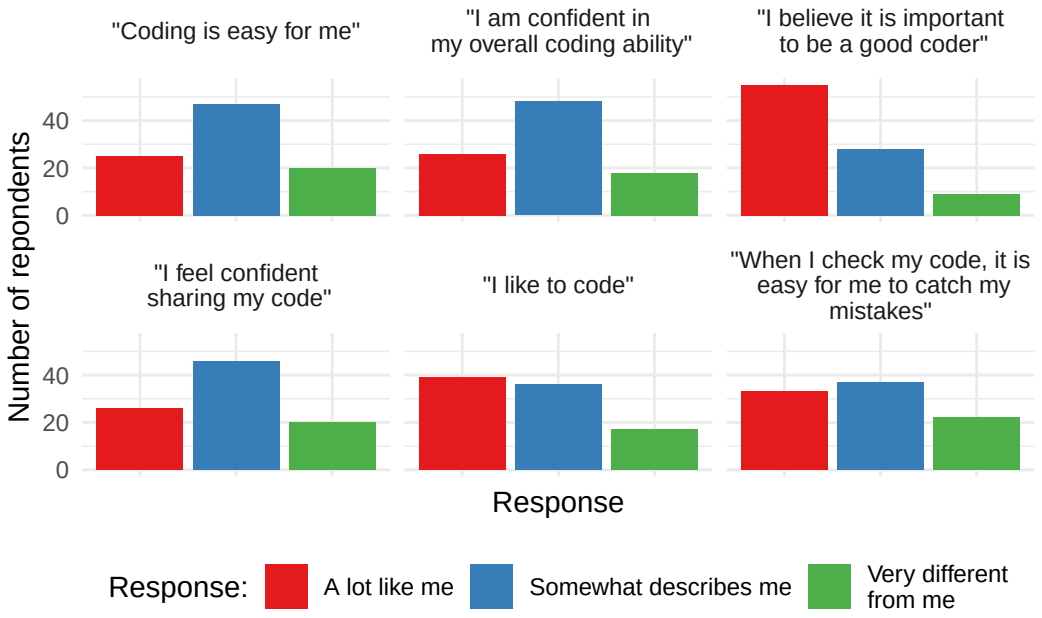
The majority of respondents were at least "somewhat familiar" with generative AI such as ChatGPT (Table 2a). Though respondents' self-perceptions around writing and coding are varied, there was strong consensus that the use of generative AI tools is appropriate within an academic setting (Table 2b). Most respondents who responded "It depends" generally found artificial intelligence tools to be appropriate, though with certain guidelines and rules governing their use.

To understand the role of LLMs in learning, students identified their usage in a more granular way by selecting various pre-defined use cases in the survey (Figure 3a). Technical questions and explaining concepts were the two top use cases among students in the course. More than half of students also used LLMs for quick questions, general knowledge, writing paper code, and checking solutions. Just less than half used it to write paper content, which could suggest that students do not feel confident using it to improve writing.

Respondents were also asked to rate the helpfulness of LLMs on various tasks assigned during the course on a 4-point scale of "Did not use" to "Very Helpful" (Figure 3b). To simplify the presentation, responses were grouped into two main categories: "Less Helpful" and "More

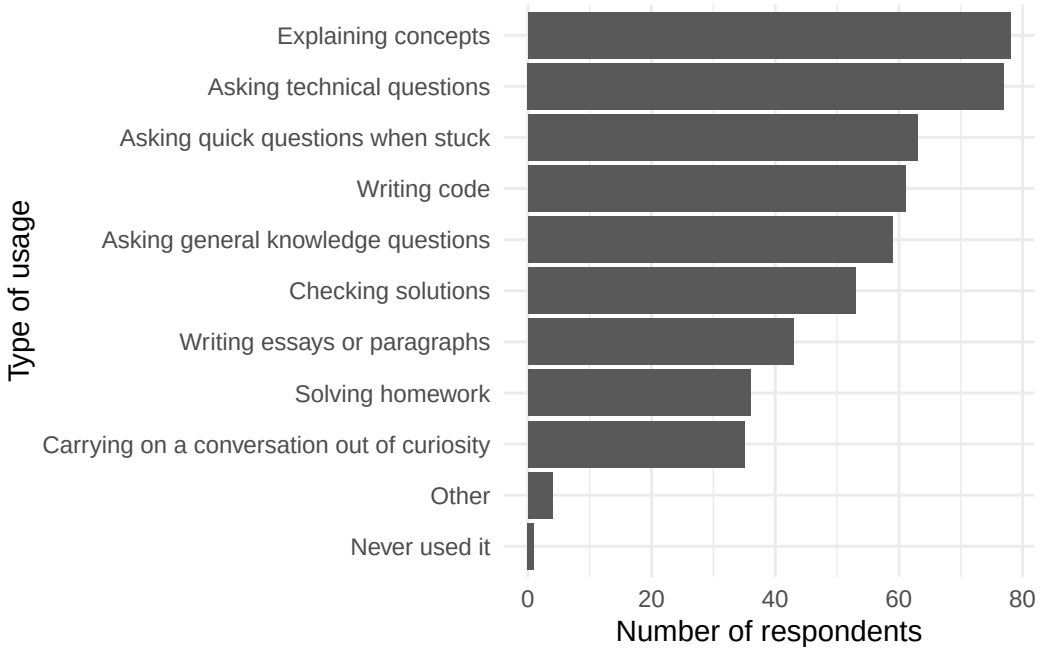


(a) "Please rate how much each statement describes you, on a scale from 'This is very different to me' to 'This is a lot like me' "

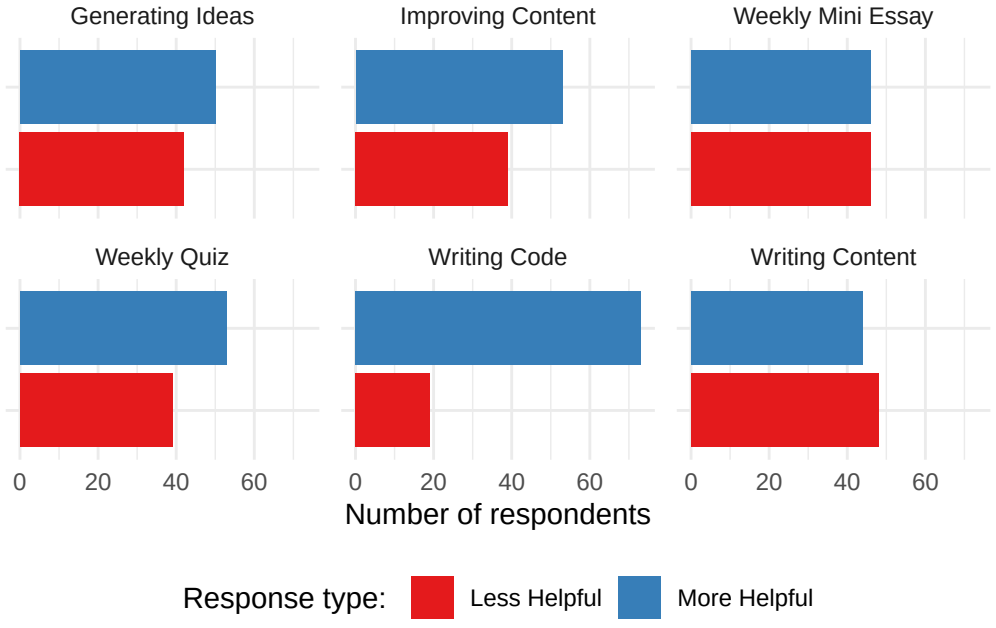


(b) "Please rate how much each statement describes you, on a scale from 'This is very different to me' to 'This is a lot like me' "

Figure 2: Self-perception of coding and writing abilities



(a) "If you have used generative AI tools such as OpenAI's ChatGPT or equivalents, in what ways have you used it (select all that apply)?"



(b) "How helpful did you find generative AI tools such as ChatGPT by OpenAI (or equivalents) for each component of STA302?"

Figure 3: Use and usefulness of generative AI

Table 2: Familiarity with, and appropriateness of, generative AI

(a) 'How familiar are you with using generative AI tools such as OpenAI's ChatGPT or equivalents?'

Extent of AI familiarity	Students
Very familiar	40
Somewhat familiar	51
Not familiar	1

(b) 'To what extent do you think using generative AI tools such as ChatGPT by OpenAI (or equivalents) is ethical and appropriate for schoolwork?'

Ethical & Appropriate for school?	Students
Appropriate	73
It depends	11
Inappropriate	8

Helpful." The "Less Helpful" category combines responses where students found the AI either "Not helpful" or did not use it for the task, while the "More Helpful" category includes responses where the LLM was considered "Somewhat helpful" or "Very helpful".

Respondents differed in terms of how they used LLMs in the course (Figure 3b). While most tasks were roughly split between respondents finding LLMs helpful or not, respondents found them most helpful in generating code. In the context of the course, this mostly meant generating R code for transforming, analyzing, and visualizing data. To a lesser extent, respondents also found LLMs to be helpful in improving the existing writing they had, while at the same time not favouring it for writing content from scratch.

Finally, one question asked students to elaborate on whether they thought generative AI tools such as ChatGPT were ethical and appropriate for schoolwork. This was an open-response question. To provide a sense of the responses, we used Anthropic's Claude 3.5 Sonnet model (as at 5 August 2024) to summarize the comments and it provided:

> Many students view AI as a helpful supplementary tool, comparing it to resources like Google or calculators. They believe it can aid in understanding concepts, debugging code, brainstorming ideas, and saving time on routine tasks. However, there's a consensus that AI should not be used to complete entire assignments or replace original thinking. Students emphasize the importance of using AI ethically, citing it when appropriate, and not relying on it exclusively. Some argue that learning to use AI effectively is a valuable skill for future careers. Concerns raised include the potential for plagiarism, the risk of hindering critical thinking skills, and the possibility of receiving incorrect information. Overall, most students support the responsible use of AI in education, with proper guidelines and transparency,

while recognizing the need to maintain academic integrity and develop independent learning skills.

2.3 LLM usage and final paper marks

Two other components were merged with the survey responses: self-reported LLM usage on the final paper, and final paper mark.

Students were encouraged to use LLMs to complete their papers. Each paper required the students to disclose their usage through a statement in the GitHub repository README for the paper. Even students who did not use generative AI at all were required to state this in the README. For students who did use generative AI, there was an additional requirement, where possible, that they save the logs of their usage in a txt file which was also included in their GitHub repository.

Those README statements were gathered and parsed using OpenAI's ChatGPT 4o model (as at 26 July). The following prompt was used:

> The following statement is about to what extent LLMs were used by a student. Please characterize it as one of: "None", "Minimal", "Somewhat", "Extensive", "Unsure". Respond with only one of those options.

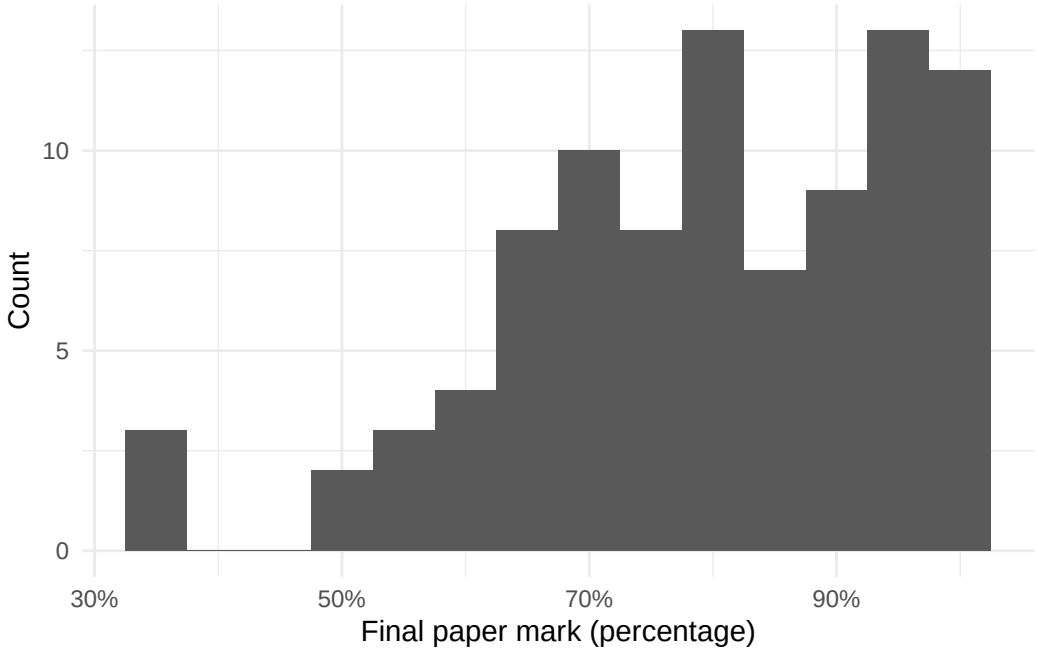
All classifications were then manually checked for reasonableness.

We find a varied extent of self-reported LLM usage (Table 3). 41 respondents were classified as having made extensive use of LLMs, while 28 were classified as having made somewhat use. 31 respondents were classified as having made minimal or no use of LLMs. In the analysis dataset we combine those two classifications because only 8 respondents were classified as having minimal usage.

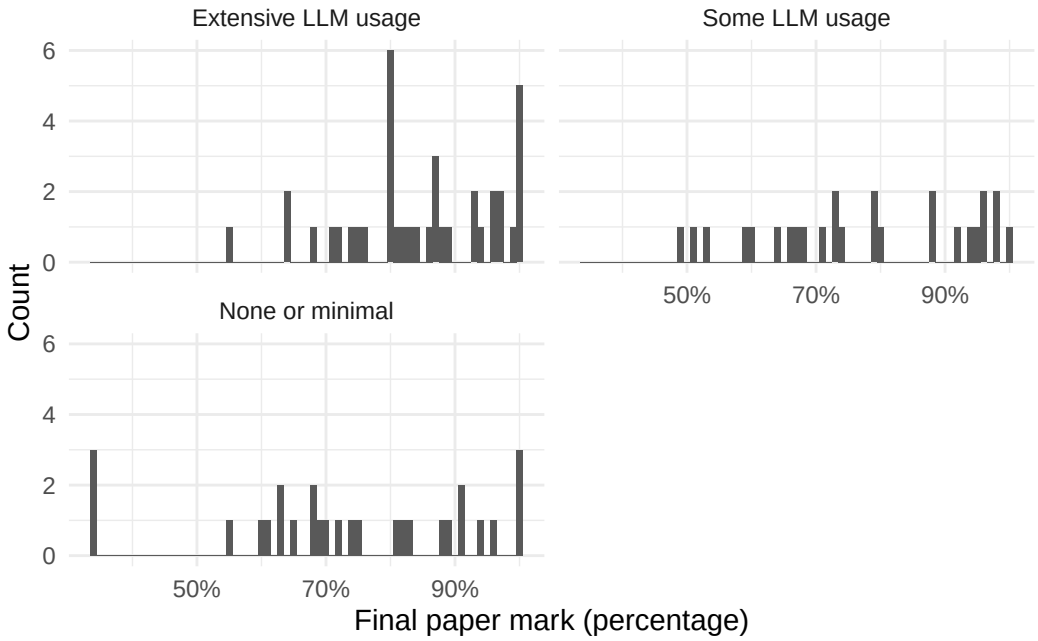
Table 3: Self-reported LLM usage in final paper

Self-reported LLM usage	Number
Extensive	38
Somewhat	26
None or minimal	28

The third, and final, component is the mark, in percentages, on the final paper (Figure 4a). The overall mean was 78% and standard deviation was 17 percentage points. However there was considerable differences by the extent of LLM usage (Table 4; Figure 4b).



(a) Overall



(b) By self-reported LLM usage

Figure 4: Distribution of marks on final paper (in percentages)

Table 4: Self-reported LLM usage in final paper and final paper mark

Self-reported LLM usage	Mean	Std dev
Extensive	0.85	0.12
Somewhat	0.77	0.16
None or minimal	0.74	0.19

3 Model

The goal of our modelling strategy is to better understand how a respondents' result on their final paper associates with their self-reported LLM usage. Students received their result on the final paper as a proportion (a 0-100% mark), so we use zero-one-inflated beta regression to best fit the distribution in this interval. We use this type of regression given the ability of the beta distribution to effectively describe continuous and skewed or heteroskedastic datasets, which we can see in the asymmetrical histogram of the distribution of marks.

Transforming the regression with a logit function as the link function ensures that the predicted values stay between zero and one, respecting the boundaries of the data. Some students got full marks for the paper, so we use the zero-one-inflated beta model.

Here we briefly describe the model that we use, which follows Kurz (2023). Model diagnostics are included in Appendix B.

OLD:

Define y_i as the percentage received on the final paper. Then β_1 and β_2 are the effect of self-reported LLM usage and self-reported GPA on the final paper grade, respectively.

$$y_i \sim \text{Beta}(\mu_i, \phi) \tag{1}$$
$$logit(\mu_i) = \beta_0 + \beta_1 \times \text{LLM usage}_i + \beta_2 \times \text{GPA}_i \tag{2}$$
$$\beta_0, \beta_1, \beta_2 \sim \text{Normal}(0, 1) \tag{3}$$
$$\phi \sim \text{Gamma}(4, 0.1) \tag{4}$$

NEW:

Let Y_i represent the dependent variable "mark" for observation i. The model assumes a zero-one inflated beta distribution for Y_i, which can be represented as:

$Y_i \sim \text{Zero-One Inflated Beta}(\mu_i, \phi_i, \pi_{zoi_i}, \pi_{coi_i})$

Where:

- μ_i is the mean of the beta distribution (related to the location parameter).
- ϕ_i is the precision parameter (related to the dispersion of the beta distribution).

- π_{zoi_i} is the probability of Y_i being exactly 0 or 1 (zero-one inflation component).
- π_{coi_i} is the probability of Y_i being in the continuous part of the beta distribution (continuous component inflation).

The specific linear models for each parameter are given by:

1. **Mean (Location) Component:** $logit(\mu_i) = \beta_{0,\mu} + \beta_{1,\mu} \cdot \text{llm\_usage}_i + \beta_{2,\mu} \cdot \text{what\_is\_your\_gpa}_i$
2. **Precision Component:** $\log(\phi_i) = \beta_{0,\phi} + \beta_{1,\phi} \cdot \text{llm\_usage}_i + \beta_{2,\phi} \cdot \text{what\_is\_your\_gpa}_i$
3. **Zero-One Inflation Component:** $logit(\pi_{zoi_i}) = \beta_{0,zoi} + \beta_{1,zoi} \cdot \text{llm\_usage}_i + \beta_{2,zoi} \cdot \text{what\_is\_your\_gpa}_i$
4. **Continuous Outcome Inflation (coi) Component:** $logit(\pi_{coi_i}) = \beta_{0,coi}$

In this model:

- $\beta_{0,\mu}, \beta_{0,\phi}, \beta_{0,zoi}$, and $\beta_{0,coi}$ are the intercepts for the respective components.
- $\beta_{1,\mu}, \beta_{1,\phi}, \beta_{1,zoi}$ represent the coefficients for the effect of "llm_usage" on the respective components.
- $\beta_{2,\mu}, \beta_{2,\phi}, \beta_{2,zoi}$ represent the coefficients for the effect of "what_is_your_gpa" on the respective components.
- The term $logit(\cdot)$ denotes the logit link function, commonly used for modeling probabilities.

This notation captures the model's structure as defined in the `brms` code, with each parameter modeled as a function of the predictors "llm_usage" and "what_is_your_gpa", except for the continuous outcome inflation (coi) component, which is modeled with just an intercept.

We estimate the model in `R` (R Core Team 2023) as well, using `brms` (Bürkner 2017).

The expected relationship between LLM usage and final mark is unclear. It could be that stronger students used LLMs in a more sophisticated way or that they did not need to use them. However, the relationship between GPA and final mark is expected to be positive.

NM Comments (I'm not familiar with 0-1 inflated beta regression so I'll leave these as comments to be considered)

- The text refers to a zero-one inflated beta, but the code chunk below is for a one-inflated model (linked to the zero-or-one inflated beta referenced, if I understand correctly). I think the text should be updated
- Is the Kurz (2023) reference correct? It looks like this is for standard beta regression, so perhaps an additional reference to one-inflated beta regression would be helpful (e.g. Ospina 2012)

- Regarding π_{zoi} and π_{coi} - based on the descriptions it seems like these should sum to 1, but if that's the case it's not clear to me why we would model both. I suspect I'm missing something here, but tweaking the descriptions could help
- llm_usage is in the model as a categorical variable, do we want to make that explicit in the model descriptions?
- Related to the above, perhaps it makes more sense to have "None" as the reference level for llm_usage instead of "Extensive"
- I typically prefer not to have raw variable names in the body of a paper, so if there's a way to do that without being to cumbersome that could be good (perhaps a personal preference though)
- Similarly, in Table 5 is it possible to use the mathematical notation instead of the variable names?
- In Figure 5 is it possible to have different symbols for the two models to make it more readable in grayscale?
- [Related to section 5.1]: There are references to how students used LLMs (e.g. to write code/text, as an alternative search engine), and I'm just curious where this information comes from. Were the LLM usage statements submitted and analysed, or is this more annecdotal? Either way, I think a bit more context about where it came from would help give the reader context.

4 Results

Our results are summarized in Table 5 and Figure 5. We especially draw on `modelsummary` (Arel-Bundock 2022).

Table 5 and Figure 5 present the results from two models estimating the association between LLM usage, GPA and final paper mark. The first model includes only LLM usage, while the second model also includes self-reported GPA.

LLM usage estimates are in relation to "Extensive" LLM usage. In the base model, the coefficients suggest that less usage of LLMs was associated with lower scores. The second model retains this slight association, but with slightly smaller coefficients. For instance, the negative impact of "b_llm_usageSomewhat" is slightly reduced from -0.35 to -0.29, indicating that part of the effect observed in the base model might be explained by the GPA variable.

Including self-reported GPA in the second model does considerably change the changes the interpretation of some coefficients. For instance, intercept for "b_zoi_Intercept" becomes much more negative in the GPA-included model, suggesting that the GPA significantly affects the zero-one inflation part of the model. However, the essential relationship between LLM usage and final mark is retained.

Table 5: Coefficient estimates and mean absolute deviation (MAD)

	Base model	Including self-reported GPA
b_Intercept	1.55	−1.42
	(0.14)	(0.52)
b_phi_Intercept	2.28	1.84
	(0.25)	(1.03)
b_zoi_Intercept	−1.86	−15.32
	(0.50)	(4.03)
b_coi_Intercept	2.64	2.67
	(1.22)	(1.21)
b_llm_usageSomewhat	−0.35	−0.29
	(0.24)	(0.21)
b_llm_usageNoneorminimal	−0.68	−0.63
	(0.23)	(0.19)
b_phi_llm_usageSomewhat	−0.59	−0.07
	(0.37)	(0.38)
b_phi_llm_usageNoneorminimal	−0.55	0.10
	(0.37)	(0.39)
b_zoi_llm_usageSomewhat	−1.43	−1.55
	(1.11)	(1.23)
b_zoi_llm_usageNoneorminimal	−0.27	−0.13
	(0.82)	(0.92)
b_what_is_your_gpa		1.00
		(0.17)
b_phi_what_is_your_gpa		0.17
		(0.33)
b_zoi_what_is_your_gpa		3.94
		(1.10)
Num.Obs.	92	92
R2	0.095	0.478
ELPD	14.2	39.3
ELPD s.e.	10.4	9.4
LOOIC	−28.4	−78.7
LOOIC s.e.	20.9	18.8
WAIC	−28.9	−79.5
RMSE	0.15	0.11

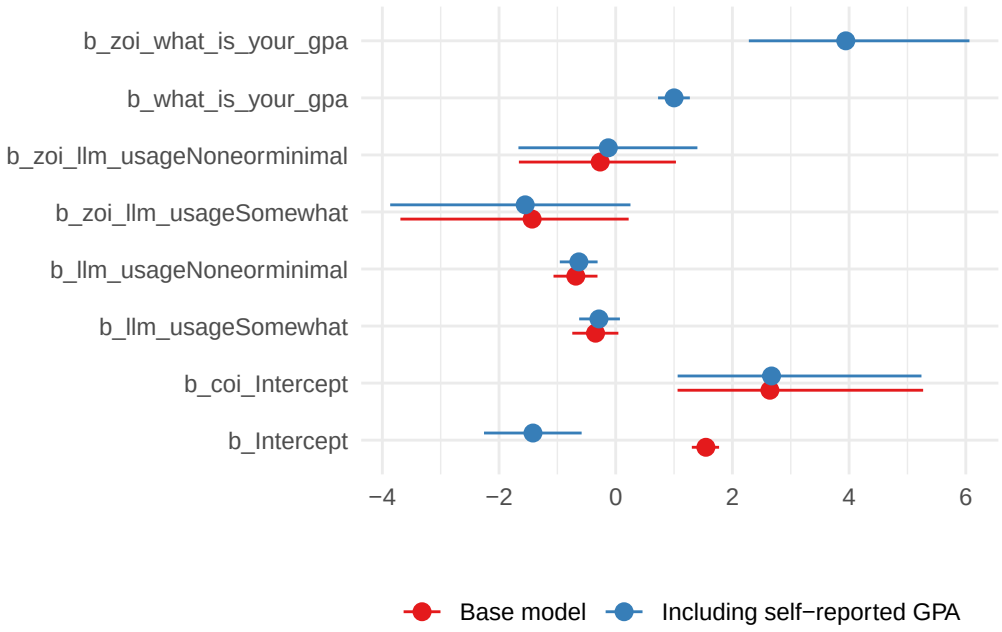


Figure 5: Selected coefficient estimates and 90 per cent credibility intervals

The overall fit of the model improves when GPA is included. There is an increase in R-squared from 0.095 to 0.478 and improvements in the ELPD and LOOIC metrics. This is also suggested by the LOO comparison between the two models (Table 6 in Appendix B).

5 Discussion

There is some very mild evidence from this observational study that LLM usage may be associated with better scores. However, this work should mostly be treated as a direction for future research in this area. More broadly, our work has three major take-aways: 1) LLMs could help enhance the trustworthiness of data science, but not without considerable futher development. 2) Making general purpose chatbots more opinionated and knowledgeable about data science would considerably help. 3) Alternative interaction approaches, beyond a chatbot interface, could be especially useful.

5.1 Impact of LLMs on trustworthy data science

We found that many students did use LLMs to help them complete a data science project closely related to what professional data scientists do. A common use case to ask the LLM to write code to make a graph. Another was to ask for some example paragraphs.

Data science code tends to be written differently than software engineering code. Functions and loops are used less, as there is often less reuse of code. The training data of these LLMs is dominated by software engineering code, and so when an LLM is asked to write code for a data science project it tends to produce code with some of those features. This is neither good nor bad in and of itself, however it can clash with the style of the code that surrounds it.

LLMs are also being used as an alternative search engine. Students often asked, for instance, how to modify a graph to add a vertical line, or similar. Here LLMs tended to be especially useful and able to answer questions. In the past students may have used StackOverflow or similar to find their answers, but the nature of the chat response likely provided a faster answer.

Finally, there is the question of whether the nature of the data science has fundamentally changed or improved. For now, there is no evidence of that. It was clear when a student had just used LLM-generated text without modification, and it was often clear when large amounts of code had been written by an LLM. But it was not obvious when small amounts of either had been used, or when LLM-generated content had just been used as an initial draft.

Many of these issues were alluded to by Bommasani et al. (2021), in the context of data science, and it is striking that they remain relevant today.

5.2 General purpose compared with opinionated specific chatbots

ChatGPT and equivalents are general-purpose chatbots. One can ask them a technical data science question and then in the same session ask for a brownie recipe! However this general-purpose nature means that they are not necessarily optimized toward one particular task (Thoppilan et al. 2022). When used for data science, this means that sometimes the functions or libraries recommended may not exist, code may not run, or out-dated approaches may be used. Frontier models make things up and are unnervingly confident, even when wrong (OpenAI et al. 2023, 59), which can create special issues for learners.

General-purpose chatbots being used for data science are fine for experienced data scientists who have an established sense of when something might be not ideal or even wrong. But it is problematic for novices who may not know, can lack the confidence to push back, and possibly, may learn the wrong approach.

Retrieval-augmented generation (RAG) means providing various sources, for instance books or manuals, which the LLM uses to base its response on. Fine-tuning means providing the previously trained LLM with a large number of example input and outputs, which again guide the response from the LLM (Raffel et al. 2019). A chatbot-based interface, focused on the needs of data science beginners, could use RAG and/or fine-tuning to considerably improve the quality of its responses. The trade-off would be a reduction in the general-purpose usability of the chatbot, but in the context of this use, such a trade-off is likely acceptable.

5.3 Changed LLM-based interfaces

LLMs existed before ChatGPT, but it was the launch of ChatGPT that brought them into the mainstream. The underlying model was not considerably different to others and the main difference that ChatGPT brought was the chat interaction. These are useful, but there is considerable opportunity to develop additional ways of interacting with LLMs. This is particularly relevant for enhancing trustworthy data science.

Search-based LLM tools like Elicit and Perplexity.ai differ to ChatGPT in that they focus on search. In contrast to Google, which typically provides a list of websites related to a particular search, Perplexity.ai attempts to provide an answer. In contrast to ChatGPT, search-based LLM tools focus on answering a question rather than engaging in discussion.

Students in this class were required to use GitHub to host their papers, as well as supporting code and data. GitHub Issues and Pull Requests were used for peer review and submission occurred based on links to GitHub repositories. Trustworthy data science could be considerably enhanced by establishing something analogous to continuous integration and a suite of tests, but for data science. The one-off nature of much data science code would mean that LLMs could be especially useful in this proposed infrastructure stack; for instance, when a data scientist commits their code or writing, a suite of static tests could run, for instance to check that the data are being read in, that classes are appropriate, etc. But additionally LLMs could be used to check for possible improvements in writing, and propose context-specific improvements in code. For instance, to make sure that what is being described in the paper is actually what is happening in the code.

Finally, one exciting area of current development in LLMs is tool use. One early example of this was enabling LLMs to use search engines such as Google. As LLMs can produce writing, and writing is the input to a Google Search, enabling this tool broadened the types of queries that LLMs could respond effectively to. Tool use, focused on data science, exists. For instance, LLMs can use R and Python to respond to queries about a dataset. However, further development could be useful. For instance, when a user wrote code in an IDE to import some data, a tool-enhanced LLM could notice and automatically write code that establishes a Pydantic-based validation model for that data. Similarly, consider an LLM in an IDE context where a data scientist wrote out the model, in statistical notation, that they were interested in estimating, and the LLM was able write the model in Stan, run it, save the estimates, and add a summary table to the paper, all automatically and in the background. Similar work has occured in other contexts and found to bring substantial benefits (Chen et al. 2021).

5.4 Weaknesses and next steps

There are several substantial weaknesses of this study. The foundational one is that we used observational data, much of which was self-reported. There may be selection bias present in terms of who used LLMs, who reported their LLM usage truthfully, and even who remained

in the class. A different design, specifically a randomized controlled trial (RCT) would deal with many of these issues, although likely at some cost.

Regression reports average estimates over the full dataset. However, many earlier studies found distributional effects, with low-performing individuals benefiting more than high-performing individuals. Again, a change in design toward an RCT could enable the exploration of this question. Stratification of our dataset would result in small sample size, but nonetheless our data does provide some limited suggestive evidence that this effect may be present in data science. For instance, looking at the 28 students who received an A+ for the final paper, 13 of them had extensive LLM usage, whereas looking at the same number of worst performing students finds that only six of them had extensive LLM usage.

Along these lines, we have considered LLM usage for code and writing as equivalent, but they should actually have different impacts depending on student backgrounds. Distinguishing between native and second-language English speakers, and then focusing on differential LLM usage, could have added a great deal of nuance to the analysis.

Finally, we only considered one outcome measure, namely grade on the final paper. Each paper required a considerable amount of time for the student to produce. If an LLM was found to reduce the time taken to produce a paper, without any reduction in quality, then that would be a similarly useful outcome.

Despite these shortcomings, our work clearly identifies a need for further research examining how LLMs can be used to develop a more trustworthy data science.

Appendix

A Survey questions

1. After carefully reading the informed consent document, please indicate below whether you consent to have your anonymized responses included in the research study?

    - Yes, I authorize the use of the data collected about me for the STA302 course survey to be used. I will be compensated 1% of my course grade for completing the survey.
    - No I do not want my data included in the research study, but I want to complete the survey. I will be compensated 1% of my course grade for completing the survey.
    - I do not want to complete this survey. I realize that I am forfeiting the corresponding course credit.

2. What is your full name on Quercus?
3. What is your Student ID?
4. What year are you?
5. What is your specialization?
6. What is/are your major/s?
7. What is/are your minor/s?
8. What is your GPA?
9. Please rate how much each statement describes you, on a scale from "This is very different to me" to "This is a lot like me" ["This is very different to me"; "This somewhat describes me"; "This is a lot like me"]

    - Writing is easy for me
    - I like to write
    - I believe it is important to be a good writer.
    - When I edit it is easy for me to catch my mistakes.
    - I feel confident sharing my writing.
    - I am confident in my overall writing ability.

10. Please rate how much each statement describes you, on a scale from "This is very different to me" to "This is a lot like me". When answering, please consider whichever programming language you are most familiar with. ["This is very different to me"; "This somewhat describes me"; "This is a lot like me"]

    - Coding is easy for me
    - I like to code
    - I believe it is important to be a good coder.
    - When I check my code it is easy for me to catch my mistakes.
    - I feel confident sharing my code.
    - I am confident in my overall coding ability.

11. How familiar are you with using generative AI tools such as OpenAI's ChatGPT or equivalents?

    - Very familiar
    - Somewhat familiar
    - Not familiar
    - Other

12. Have you used any generative AI tools such as OpenAI's ChatGPT or equivalents for any reason (personal or educational)?

    - Yes
    - No
    - Other

13. If you have used generative AI tools such as OpenAI's ChatGPT or equivalents, in what ways have you used it (select all that apply)?

    - Asking technical questions
    - Carrying on a conversation out of curiosity
    - Asking general knowledge questions
    - Solving homework
    - Checking solutions
    - Asking quick questions when stuck
    - Explaining concepts
    - Writing essays or paragraphs
    - Writing code
    - Never used it
    - Other

14. To what extent do you think using generative AI tools such as ChatGPT by OpenAI (or equivalents) is ethical and appropriate for schoolwork?

    - Appropriate
    - Inappropriate
    - Other

15. Please elaborate on your answer above.
16. Did you use any generative AI tools such as OpenAI's ChatGPT or equivalents for STA302?

    - Yes
    - No
    - Other

17. How helpful did you find generative AI tools such as ChatGPT by OpenAI (or equivalents) for each component of STA302? ["Not helpful"; "Somewhat helpful"; "Very helpful"; "I did not use generative AI for this component"]

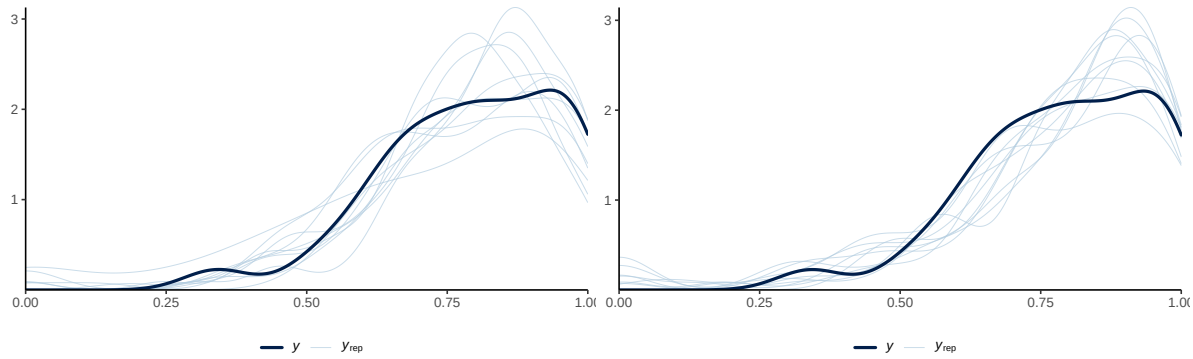
- Weekly quiz
- Weekly mini-essay
- Papers: Generating ideas
- Papers: Writing code
- Papers: Writing content
- Papers: Improving content

18. What is your recommendation for how generative AI tools such as ChatGPT by OpenAI (or equivalents) should be used in the course in future?
19. (Optional) Any other comments?

B Model details

We use `bayesplot` (Gabry and Mahr 2024) and `loo` (Vehtari et al. 2024) conduct posterior predictive checks and evaluate model diagnostics.

B.1 Posterior predictive check



(a) Base model

(b) Including self-reported GPA

Figure 6: Posterior predictive checking

Table 6: Model comparison

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
fit2	0.000	0.000	39.345	9.385	11.538	2.036	-78.690	18.769
fit1	-25.125	6.548	14.220	10.426	8.858	1.541	-28.441	20.852

B.2 Diagnostics

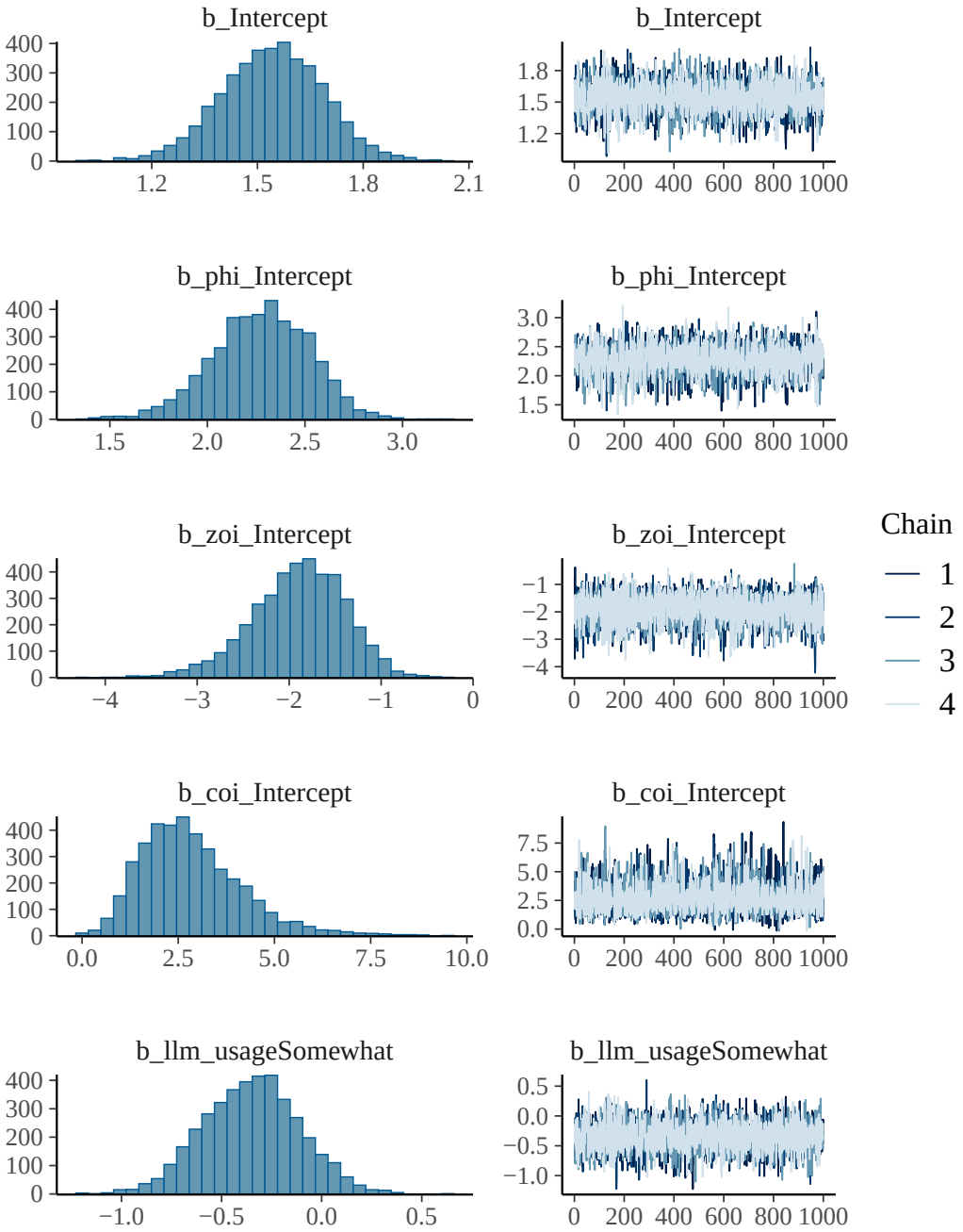


Figure 7: Base model diagnostics

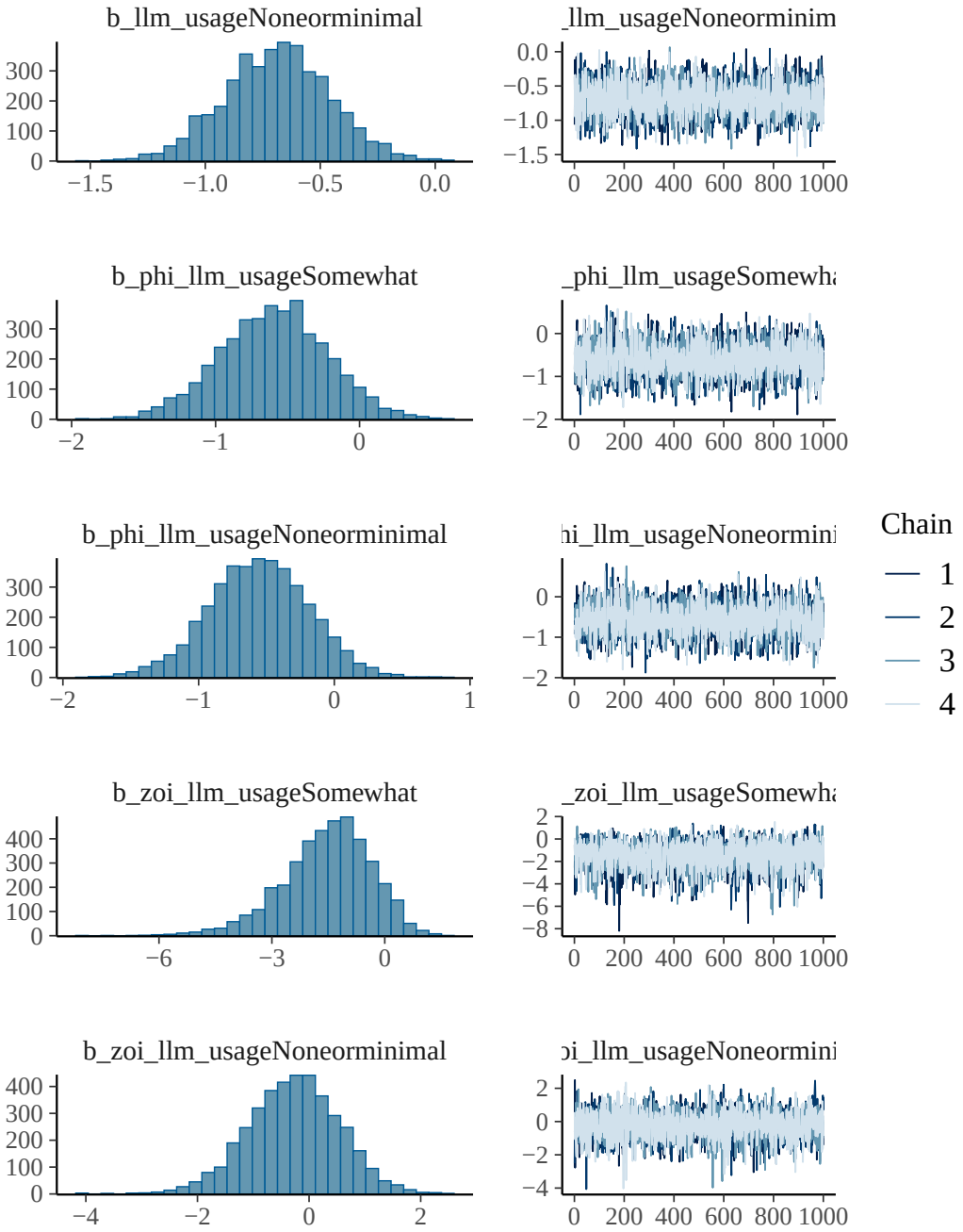


Figure 8: Base model diagnostics

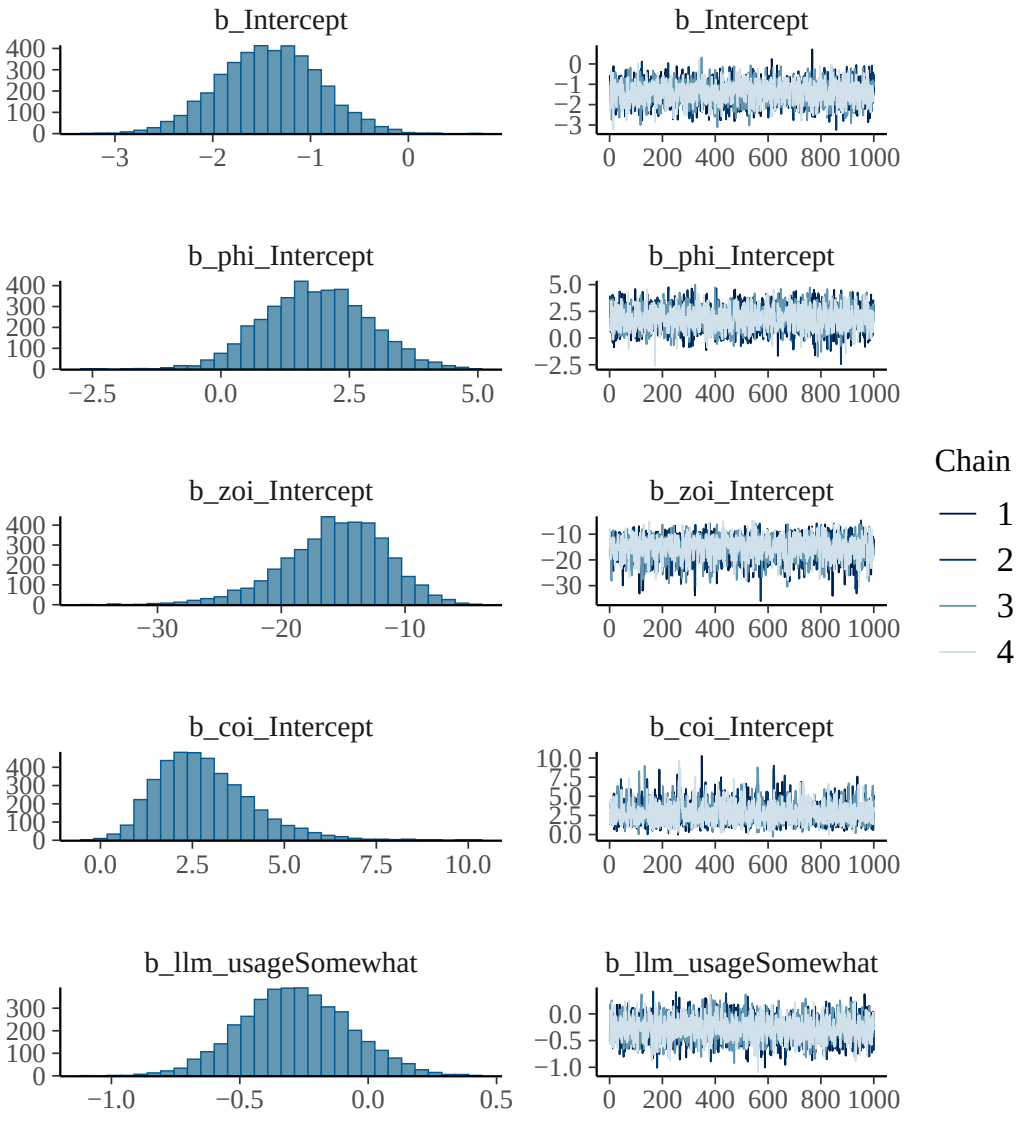


Figure 9: Model including self-reported GPA diagnostics

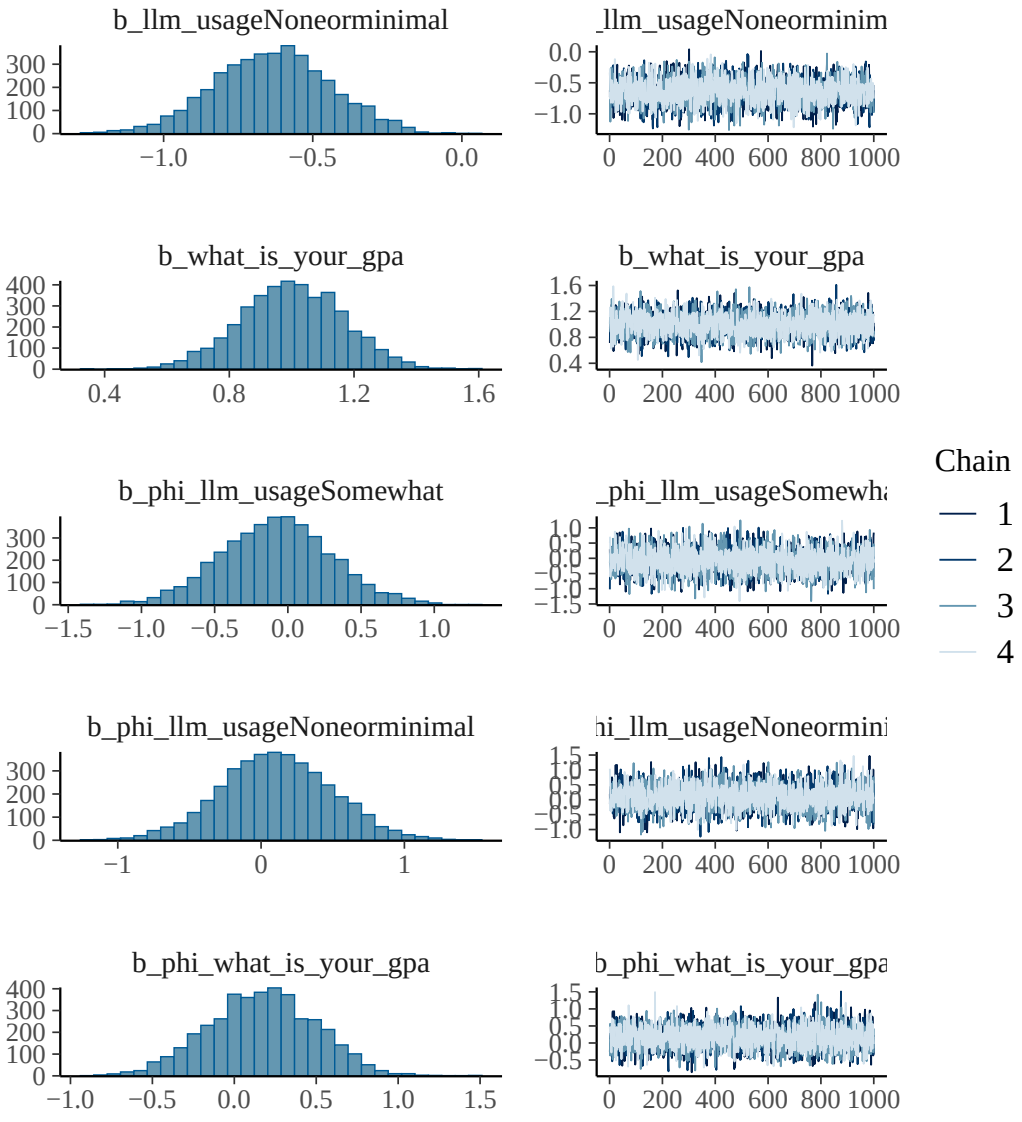


Figure 10: Model including self-reported GPA diagnostics

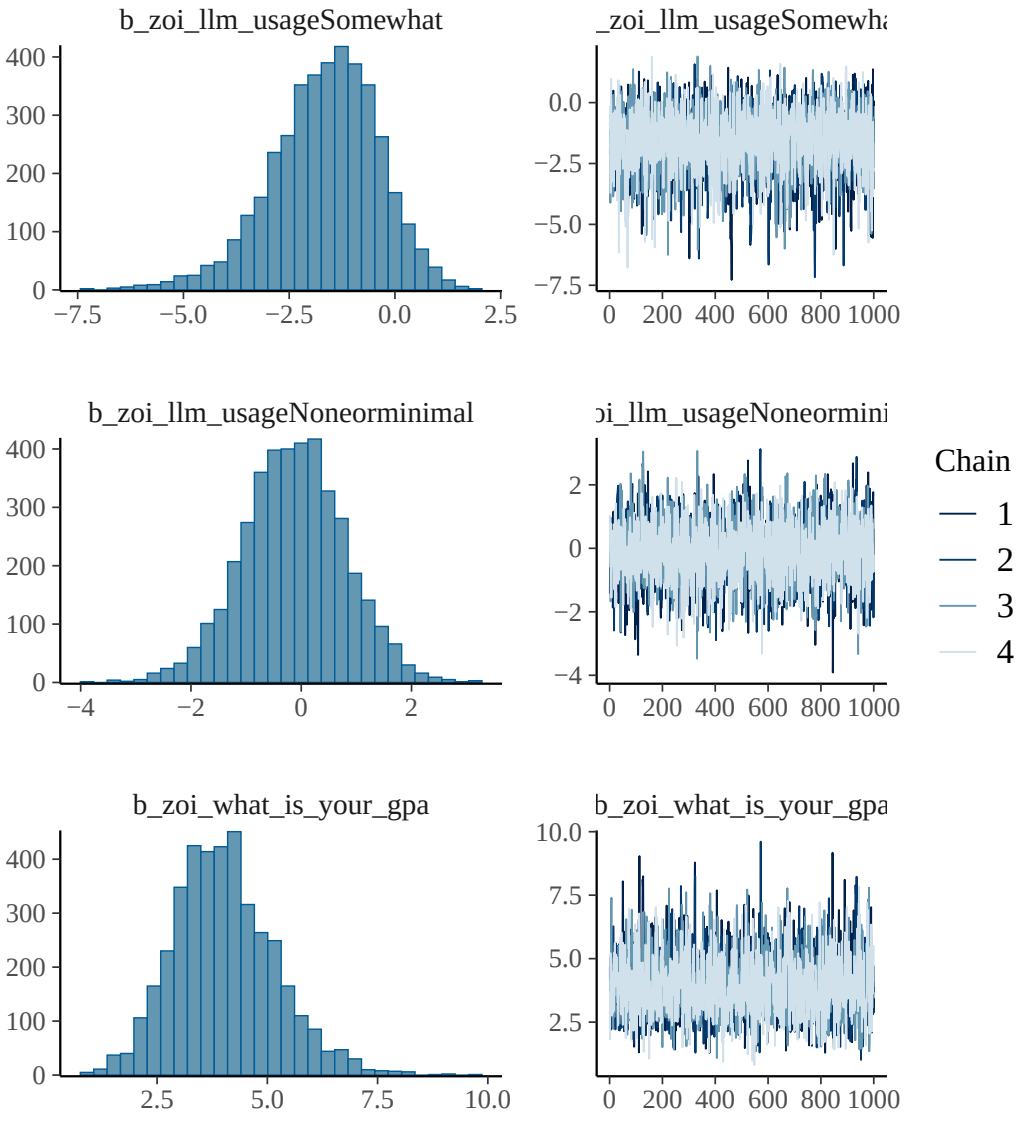


Figure 11: Model including self-reported GPA diagnostics

References

Adhikari, Ani, John DeNero, and Michael I. Jordan. 2021. "Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley." *Harvard Data Science Review* 3 (2). https://doi.org/10.1162/99608f92.cb0fa8d2.

Afzal, Samra, Shazia Zamir, and Muhammad Asghar Ali. 2024. "Tailoring Shadow Education: Leveraging ChatGPT to Transform Private Tutoring Landscape to Optimized Performance of Students." *Journal of Development and Social Sciences* 5 (2): 313–23. https://doi.org/10.47205/jdss.2024(5-II)30.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Baidoo-Anu, David, and Leticia Owusu Ansah. 2023. "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning." *SSRN Electronic Journal*. https://api.semanticscholar.org/CorpusID:256347543.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. "On the Opportunities and Risks of Foundation Models." arXiv. https://doi.org/10.48550/ARXIV.2108.07258.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

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.

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. "Evaluating Large Language Models Trained on Code." arXiv. https://doi.org/10.48550/ARXIV.2107.03374.

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.

Eke, Damian Okaibedi. 2023. "ChatGPT and the Rise of Generative AI: Threat to Academic Integrity?" *Journal of Responsible Technology* 13 (April). https://doi.org/10.1016/j.jrt.2023.100060.

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.

Fulgencio, Sánchez-Vera. 2024. "Developing Effective Educational Chatbots with GPT: Insights from a Pilot Study in a University Subject." *Trends in Higher Education* 3 (1): 155–68. https://doi.org/10.3390/higheredu3010009.

Gabry, Jonah, and Tristan Mahr. 2024. "bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

Gibbs, Alison L., and Nathan Taback. 2021. "The Building Blocks of Statistical Education in the Data Science Ecosystem." *Harvard Data Science Review* 3 (2). https://doi.org/10.1162/99608f92.8bb28793.

Horton, Diane, David Liu, Sheila A. McIlraith, Steven Coyne, and Nina Wang. 2024. "Do Embedded Ethics Modules Have Impact Beyond the Classroom?" In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education v.1*. SIGCSE 2024. ACM. https://doi.org/10.1145/3626252.3630834.

Kurz, Solomon. 2023. "Causal Inference with Beta Regression," June. https://solomonkurz.netlify.app/blog/2023-06-25-causal-inference-with-beta-regression/.

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). https://doi.org/10.1162/99608f92.9f0adc39.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. "GPT-4 Technical Report." arXiv. https://doi.org/10.48550/ARXIV.2303.08774.

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/.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." arXiv. https://doi.org/10.48550/ARXIV.1910.10683.

Rochwerg, Benny. 2024. "Evidence of Racial Profiling by the Austin Police Department in 2020." https://github.com/bennyrochwerg/profiling/blob/main/paper/paper.pdf.

Su, Emily. 2024. "Characteristics of Top Songs Has Changed from Pandemic Brain: An analysis of songs on Billboard's Year-End Hot 100 singles list (2014 to 2023)." https://github.com/moonsdust/top-songs/blob/main/paper/paper.pdf.

Thiga, Moses Mwangi. 2024. "Generative AI and the Development of Critical Thinking Skills." *Iconic Research And Engineering Journals* 7: 83–90.

Thoppilan, Romal, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, et al. 2022. "LaMDA: Language Models for Dialog Applications." arXiv. https://doi.org/10.48550/ARXIV.2201.08239.

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). https://doi.org/10.1162/99608f92.bff007ab.

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.

Vehtari, Aki, Jonah Gabry, Måns Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman. 2024. "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." https://mc-stan.org/loo/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

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.

Wickham, Hadley, and Jennifer Bryan. 2023. *readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Yu, Hannah. 2024. "Fake News vs Fox News: The Influence of Media Preferences on Voting Behavior in the 2020 U.S. Presidential Election Among Party Voters." https://github.com/hannahyu07/Fox-News/blob/main/paper/Fake_News_vs_Fox_News.pdf.