Self-reported LLM usage and outcomes on a data science project: Evidence from two undergraduate data science courses*

REDACTED REDACTED REDACTED

April 7, 2025

To help understand the effect of Large Language Models (LLMs) on data science practice we examine the extent to which self-reported LLM usage is correlated with the mark that a student received on a final paper in a classroom data science setting. We find some mild evidence from this observational study that LLM usage may be associated with better scores, especially for students who do not natively speak English. Comparing self-reported usage, there was a considerable increase between the class that occurred in January-April 2024 where 41 per cent of students self-reported extensive LLM usage and the class that occurred in September-December 2024 where 69 per cent reported extensive LLM usage. Despite the classroom setting used for evaluation, the task 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 in both the classroom and the workplace.

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 including ongoing education, adherence to professional norms 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 accountability, and critically consider the broader impact of data-driven decisions. Recent advances, such as the

*Code and data are available at: REDACTED. We thank attendees at JSM 2024 for helpful suggestions. This research was approved by the University of REDACTED's Research Ethics Board (Protocol 47725). Comments: REDACTED.

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 used 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; Bastani et al. 2024).

The tasks involved in a trustworthy data science workflow can be generally broken down into a few 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 follows 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 per cent 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 per cent increase in delivery speed of business tasks, most of which involved some writing, as well as a 40 per cent increase in human-rated performance on those tasks. Like computer programming, these productivity increases were most pronounced for those with below average performance, with their output increasing by 43 per cent.

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. Lehmann, Cornelius, and Sting (2025) found the effect of LLMs on learning was nuanced and depended on the way the LLM was used and the level of prior knowledge. At the high school level Lazar et al. (2023) found that LLMs could help creativity, provide academic support when teachers were unavailable, and model certain types of writing well. But teachers were found to have concerns about over-reliance, academic integrity, 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. Only 11 per cent "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.

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 two upper-year undergraduate data science courses – one in Winter 2024 (January to April) and another in Fall 2024 (September to December) – and whether students who used LLMs tended to have higher scores than those who did not. We focus on the association between student academic performance and their LLM usage. Specifically, we examine the relationship between the mark a student got on a final paper 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 third-year undergraduate data science courses at the University of REDACTED. 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.

We find some mild association between the mark that a respondent received and their self-reported LLM usage. And this is especially the case for respondents that were not native English speakers. More concretely, there was a considerable increase in the number of respondents who self-reported extensive LLM usage and also those who considered LLM usage to be appropriate and ethical for university, when we compared the results in the class at the start of 2024 with the class at the end of 2024. Taken together it is clear that considerable additional work is needed understanding the appropriate use of generative AI in statistics and data science pedagogy.

The remainder of this paper is structured as follows: Section 2 visualizes and analyzes survey data, self-reported LLM usage, and final marks. 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 students taking STA302 "Methods of Data Analysis I" in the Winter 2024 semester (January to April 2024), or students taking STA304 "Surveys, Sampling, and Observational Data" in the Fall 2024 semester (September to December 2024), at the University of REDACTED. STA302 had 275 students initially enrolled which reduced to 154 students by the end of the semester. Similarly, STA304 had 275 students initially, reduced to 196 students by the end of the semester. These reductions reflect a normal rate of attrition for undergraduate statistics courses at the University of REDACTED. Assessment in both courses 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 was done individually. By this stage, uninterested students have typically dropped the course, and students are familiar with course expectations. A typical paper submission was 10-20 pages, and required 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: REDACTED; REDACTED; and REDACTED.

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 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. In STA302 formal instruction in LLM usage 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. In STA304 formal instruction was provided by instructor-demonstration of efficient LLM usage for coding and writing.

Data were collected from students through an optional end-of-course survey. Appendix A details the questions asked in the survey for STA302. The STA304 survey was almost identical, with the one difference discussed later. Whether or not they consented to their data being used, all respondents received a 1 per cent increase in their final course grade for filling out the survey. Consenting responses were then matched to their final paper mark, as well as the GitHub repository for their final paper. The responses were then anonymized by removing any personal references to the students themselves including, names, emails, student numbers, GitHub links, and summarizing free-text responses.

Data cleaning and analysis was done using the `R` statistical programming language (R Core Team 2025), and the `arrow` (Richardson et al. 2024), `here` (Müller 2020), `janitor` (Firke 2023), `readxl` (Wickham and Bryan 2023), and `tidyverse` (Wickham et al. 2019) packages.

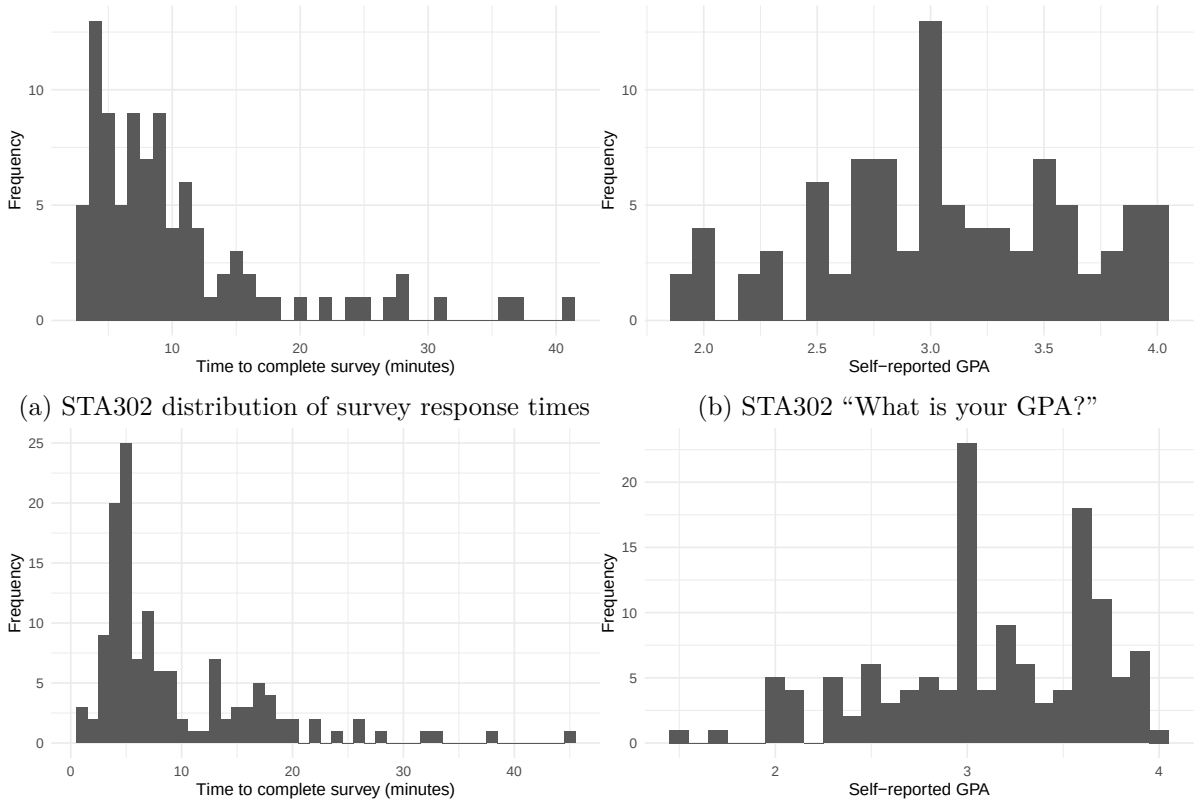
2.2 Survey data

In STA302, 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, seven of those respondents did not provide a usable GPA response.

In STA304, there were 183 responses to the survey. Of these, 153 respondents provided authorization for their data to be collected and used. Three of those respondents submitted the survey twice and their second response was removed. One student submitted their final paper too late to be of use, and so their survey was removed. Nine respondents did not include a statement on LLM usage in the README of the GitHub repository of their final paper. One respondent did not provide a usable GPA response. Finally, one respondent failed an attention check question halfway through the survey "Please select the option blue".

In STA302, all but one respondent completed the survey within 60 minutes. After that respondent was removed from the analysis the average time to complete the survey was 11 minutes, and the standard deviation was 8 minutes (Figure 1a). This left 92 respondents from STA302 that were of use and were merged based on student name.

In STA304 three students took longer than one hour to complete the survey and were removed. After those respondents were removed the average time to complete the survey was 10 minutes and the standard deviation was 8 minutes (Figure 1b). This left 131 respondents that were of use and were merged based on name.



(a) STA302 distribution of survey response times

(b) STA302 "What is your GPA?"

(c) STA304 distribution of survey response times

(d) STA304 "What is your GPA?"

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

Table 1: "What year are you?"

(a) STA302 responses (Winter 2024)

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

(b) STA304 responses (Fall 2024)

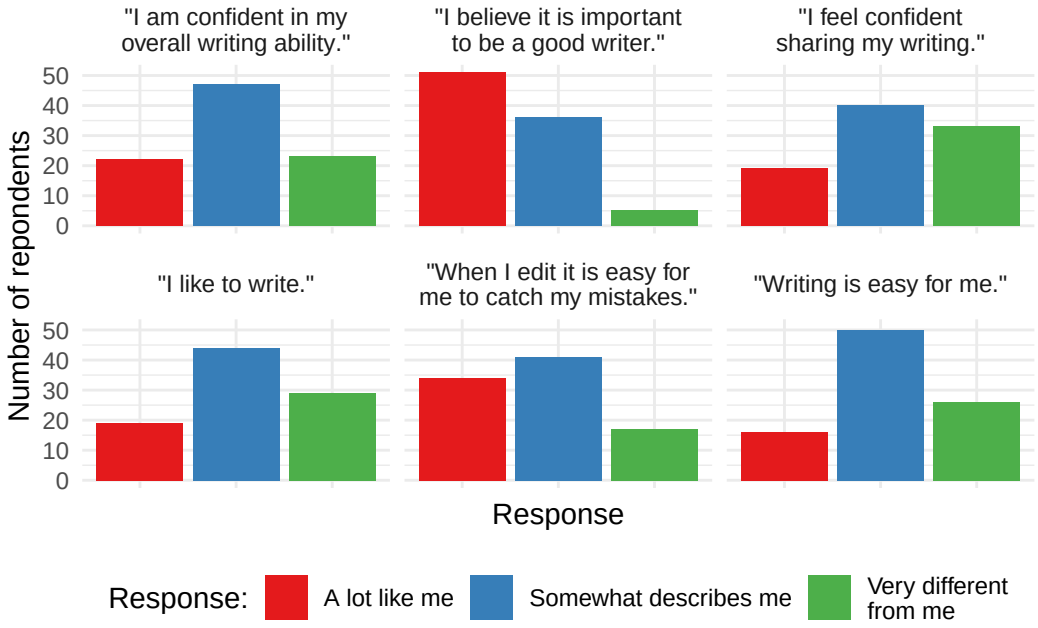
2nd	3rd	4th	5th or over
2	45	69	15

In STA302 there was a wide distribution of self-reported GPAs (Figure 1b). Most responses clustered around a B (3.0 out of 4.0), and the average was 3.06 with a standard deviation of 0.55. One factor that may have affected the range is that the course is required for programs in the Statistics, Mathematics and Computer Science Departments. While there was no reason for the respondents to not report the truth, 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. The mean and standard deviation of self-reported GPA was the same in STA304 (Figure 1d).

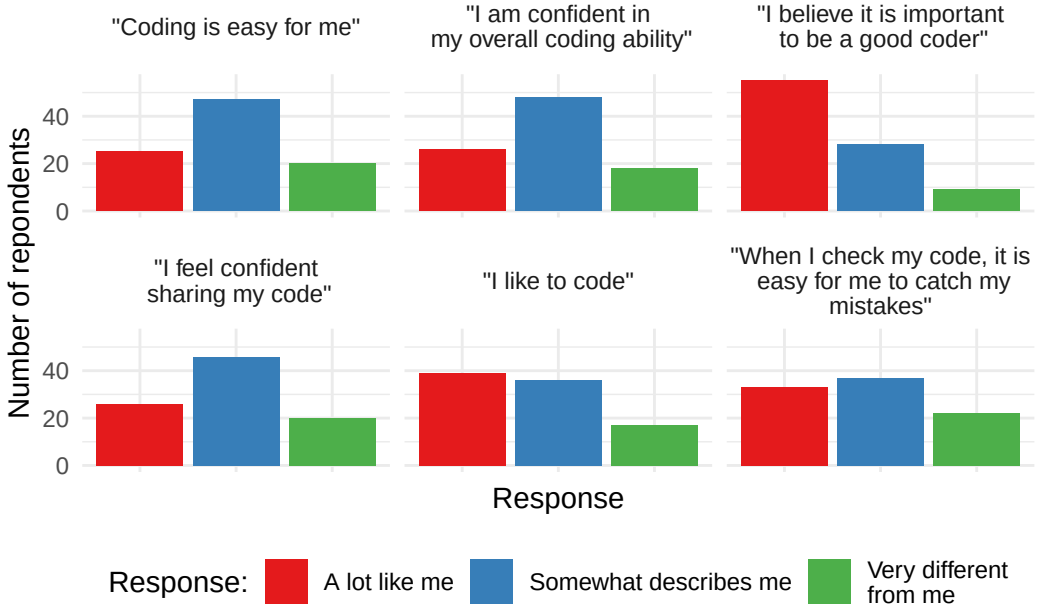
Students from a range of years took both courses, but most respondents were in their 3rd or 4th year of study (Table 1). The prerequisite two-course sequence is typically completed by students in their second year, would make it difficult to take either course earlier than their 3rd year.

Respondents from STA302 had a varied self-perception of their coding and writing abilities (Figure 2). Most respondents believed it was important to be good at writing, but many were 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 was strong consensus on the perceived importance of coding skills. However, respondents' self-assessed coding abilities and enjoyment were 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. 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



(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 in STA302 (Winter 2024)

Table 2: 'To what extent do you think using generative AI tools such as ChatGPT by OpenAI (or equivalents) is ethical and appropriate for schoolwork?'

(a) Responses from STA302 (Winter 2024)

Ethical & Appropriate for school?	Number	Percentage
Appropriate	73	79
It depends	11	12
Inappropriate	8	9

(b) Responses from STA304 (Fall 2024)

Ethical & Appropriate for school?	Number	Percentage
Appropriate	114	87
It depends	12	9
Inappropriate	5	4

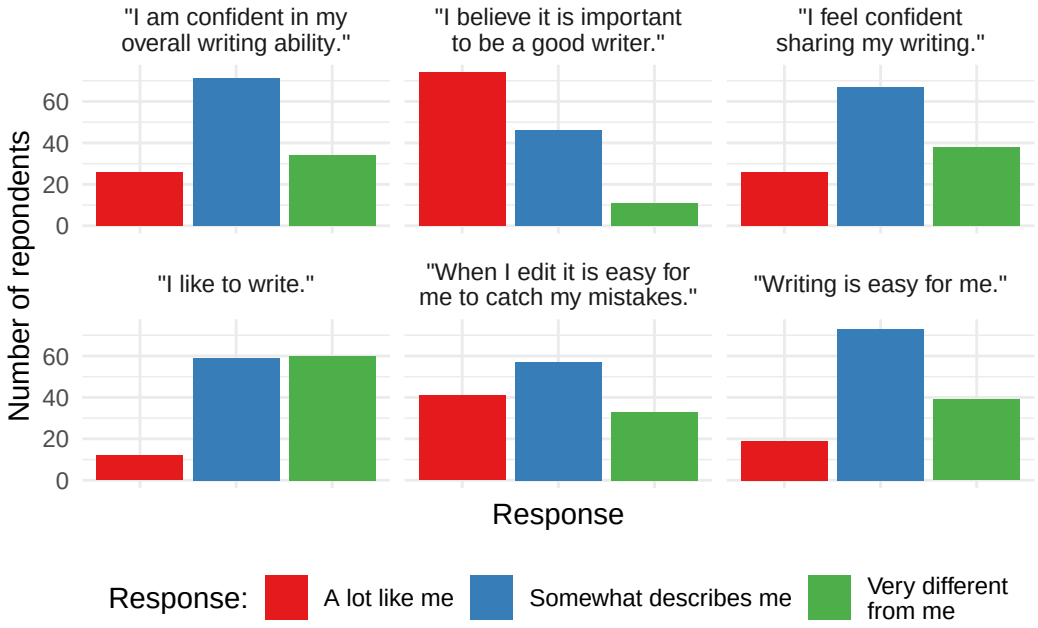
competence and comfort across various coding-related activities.

Similar results were also found in STA304 (Figure 3).

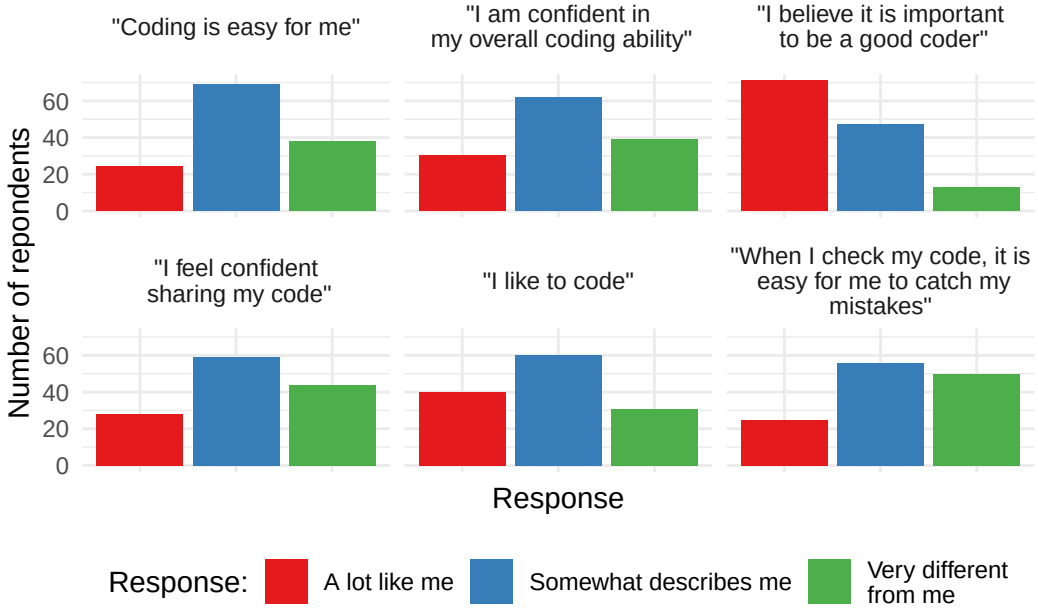
Though STA302 respondents' self-perceptions around writing and coding were varied, there was strong consensus (79 per cent) that the use of generative AI tools was appropriate within an academic setting (Table 2a). Most respondents who selected "It depends" (12 per cent) generally found in comments that artificial intelligence tools were appropriate, though with certain guidelines and rules governing their use. However, in STA304, almost all students (87 per cent) felt that it was ethical and appropriate to use generative AI for school and only 4 per cent felt that it was inappropriate. STA302 ran from January to April 2024, while STA304 ran from September to December 2024.

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 4). To simplify the presentation, responses were grouped into two main categories: "Less Helpful" and "More 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 STA302 (Figure 4a). 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 favoring it for writing content from scratch. This was similar to the results found for STA304 (Figure 4b), although there was consistently increased usage for all four aspects.

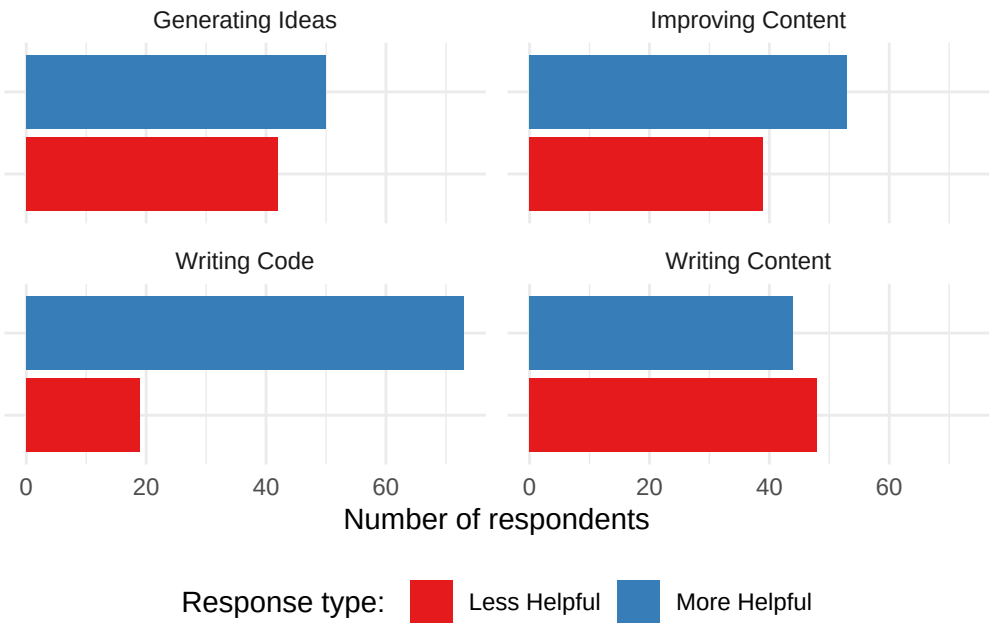


(a) "Please rate how much each statement describes you, on a scale from 'This is very different to me' to 'This is a lot like me' "

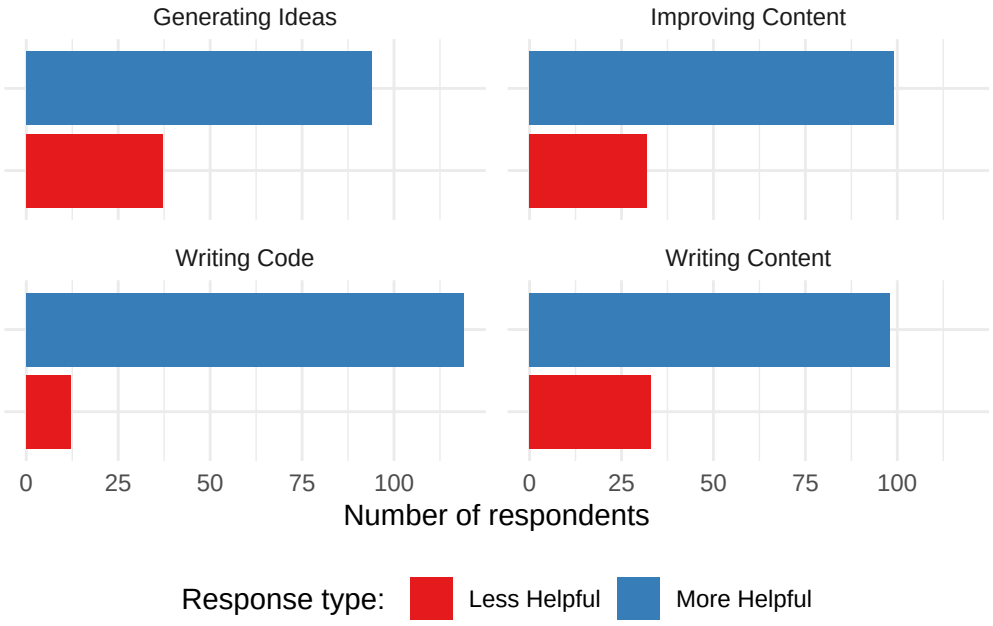


(b) "Please rate how much each statement describes you, on a scale from 'This is very different to me' to 'This is a lot like me' "

Figure 3: Self-perception of coding and writing abilities in STA304 (Fall 2024)



(a) Responses from STA302 (Winter 2024)



(b) Responses from STA304 (Fall 2024)

Figure 4: "How helpful did you find generative AI tools such as ChatGPT by OpenAI (or equivalents) for different components?"

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 from STA302, we used Anthropic's Claude 3.5 Sonnet model (as at 5 August 2024) to summarize the comments and it provided:

> Many respondents 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. Respondents 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 respondents 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.

As at 14 August 2024, the default Anthropic setting is that they do not use input data for training and so these student responses should not have entered future training datasets (Anthropic 2024).

We did a similar exercise for the responses from STA304, however using OpenAI's o3-mini model (as at 17 March 2025). It provided:

> The responses indicate a general consensus that generative AI tools, such as ChatGPT, can be ethical and appropriate for schoolwork when used as a supplement rather than a substitute for genuine effort. Many liken their use to that of calculators or grammar checkers, highlighting benefits like idea generation, code debugging, and clarifying complex concepts. However, a common caution is that overreliance on these tools—especially for tasks that require critical thinking and original work—could undermine learning. Transparency, proper citation, and maintaining a balance between AI assistance and personal input are repeatedly emphasized, ensuring that the tools aid understanding without replacing the student's own analytical process.

Finally, in STA304 only, we asked about whether the student was an international student and whether they spoke English as their first language (Table 3). It is notable that most respondents are international students (Table 3a) and did not speak English as their first language (Table 3b).

2.3 LLM usage and final paper marks

Two other components were merged with the survey responses:

Table 3: The background of respondents in STA304 only

(a) 'Are you an international student?'

International student?	Number	Percentage
No	45	34
Yes	86	66

(b) 'Is English your native language (i.e. the language that you first spoke)?'

Native English speaker?	Number	Percentage
No	97	74
Yes	34	26

1) self-reported LLM usage on the final paper, and
2) final paper mark.

Students were encouraged to use LLMs to complete their papers. But 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 text file which was also included in their GitHub repository.

Those README statements were gathered and parsed using OpenAI's ChatGPT 4o model. For the STA302 statements this was as at 26 July, 2024, while for the STA304 statements this was as at 16 March 2025. Between this time period there were underlying changes to the model despite no change in the model name. 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. In the case of STA304, there were 31 cases where the statement provided by the student in the README was vague, and the LLM had classified it at "Somewhat". Manual inspection of text files provided by the students resulted in 19 of these being re-classified as "Extensive", 7 of them being re-classified as "Somewhat", and 5 of them being re-classified as "Minimal". None of them were re-classified as "None".

We find a varied extent of self-reported LLM usage (Table 4). In STA302, 38 respondents were classified as having made extensive use of LLMs, while 26 were classified as having made somewhat use. 28 respondents were classified as having made minimal or no use of LLMs. In the analysis dataset we combine those two classifications because only a handful of respondents were classified as having minimal usage. In STA304, which occurred only six months later,

Table 4: Self-reported LLM usage in final paper

(a) Responses from STA302 (Winter 2024)

Self-reported LLM usage	Number	Percentage
Extensive	38	41
Somewhat	26	28
None or minimal	28	30

(b) Responses from STA304 (Fall 2024)

Self-reported LLM usage	Number	Percentage
Extensive	91	69
Somewhat	24	18
None or minimal	16	12

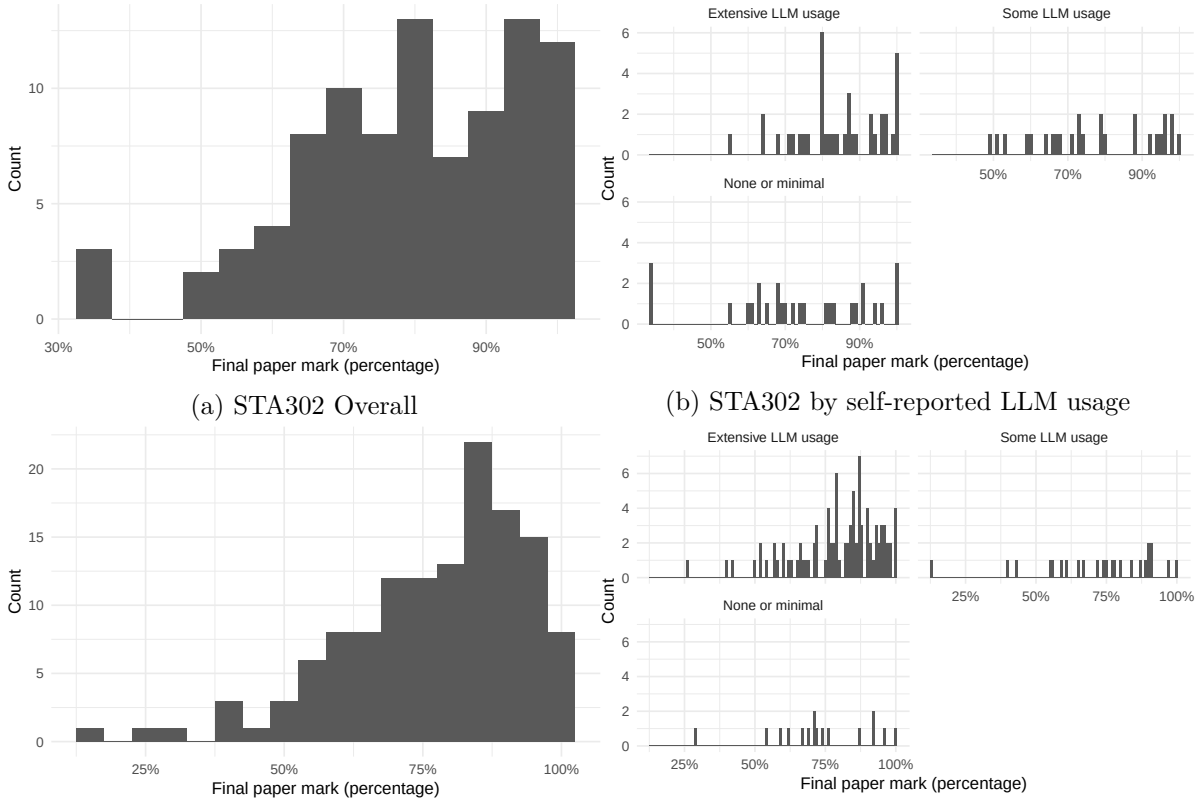
we find that a vast majority (69 per cent) of students made extensive use of LLMs and only 12 per cent reported none or minimal usage. Partly this may be students having become more comfortable with the idea of reporting LLM usage but it also likely reflects a change in actual usage. Either way, the increase in the percentage of students reporting extensive LLM usage and the decrease in the percentage reporting none or minimal usage is notable and has important implications for statistics and data science pedagogy.

The third, and final, component is the mark, in percentages, on the final paper (Figure 5). In STA302 the overall mean was 79 per cent and standard deviation was 16 percentage points (Figure 5a) while in STA304 the overall mean was 78 per cent and the standard deviation was 17 percentage points (Figure 5c). However, in both classes there were considerable differences by the extent of LLM usage (Figure 5b; Figure 5d; Table 5a; Table 5b).

3 Model

The goal of our modelling is to better understand how a respondents' result on their final paper associates with their self-reported LLM usage in our dataset. Students received their result on the final paper as a percentage between 0 per cent and 100 per cent. There were no students who got 0 per cent, but some students got 100 per cent. As such, after converting the percentage to a proportion, we use one-inflated beta regression (Ospina and Ferrari 2012). Beta regression is commonly used when data are between 0 and 1 and is focused on estimating the two shape parameters that govern the beta distribution. When there are some values that are 1, a variant – one-inflated beta regression – can be used, which is a mixture of focusing on whether a value was 1 or not, and then the usual beta distribution aspect.

The main predictor of interest, `llm_usage`, is a categorical variable with three possible values: "Extensive", "Somewhat", "None or minimal". We use "None or minimal" as the reference



(a) STA302 Overall

(b) STA302 by self-reported LLM usage

(c) STA304 Overall

(d) STA304 by self-reported LLM usage

Figure 5: Distribution of marks on final paper (in percentages)

Table 5: Self-reported LLM usage in final paper and final paper mark

(a) Responses from STA302 (Winter 2024)

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

(b) Responses from STA304 (Fall 2024)

Self-reported LLM usage	Mean	Std dev
Extensive	0.80	0.15
Somewhat	0.72	0.21
None or minimal	0.73	0.18

level and so the results are in relation to that level of self-reported LLM usage. Self-reported GPA, `what_is_your_gpa`, is a number between 1.5 and 4.0, with one decimal place.

If y_i is the proportion that student i received on the final paper, then our model is:

$$y_i \sim \text{One-Inflated Beta}\left(\mu_i, \phi_i, \pi_{\text{zoi}_i}, \pi_{\text{coi}_i}\right)$$

where:

- μ_i (mean component) is the mean of the beta distribution, which accounts for marks between zero and one,
- ϕ_i (precision component) is the precision of the beta distribution, which, again, accounts for marks between zero and one,
- π_{zoi_i} (zero-one inflation component) is a logistic regression model that looks at which of two groups a student is in: 1) the group of students getting full or zero marks, or 2) the group of students getting a mark between zero and one, and
- π_{coi_i} (continuous outcome inflation component) is used to distinguish between students that got full or zero marks (but in our case as there were no students that got zero we force this to be one).

The main model specification is:

1. Mean component: $\text{logit}(\mu_i) = \beta_{0,\mu} + \beta_{1,\mu} \cdot \text{llm\_usage}_i + \beta_{2,\mu} \cdot \text{GPA}_i$
2. Precision component: $\log(\phi_i) = \beta_{0,\phi} + \beta_{1,\phi} \cdot \text{llm\_usage}_i + \beta_{2,\phi} \cdot \text{GPA}_i$
3. Zero-one inflation component: $\text{logit}(\pi_{\text{zoi}_i}) = \beta_{0,\text{zoi}} + \beta_{1,\text{zoi}} \cdot \text{llm\_usage}_i + \beta_{2,\text{zoi}} \cdot \text{GPA}_i$
4. Continuous outcome inflation component: $\text{logit}(\pi_{\text{coi}_i}) = \beta_{0,\text{coi}}$.

We estimate the model and explore the results using the statistical programming language `R` (R Core Team 2025) and the `brms` (Bürkner 2017) package. Our code to adjust for the fact that there were no students who got zero follows Bürkner (2020) and Heiss (2021). Model diagnostics are included in Appendix B.

Finally, while our approach of using zero-inflated beta regression is a common one, another approach for dealing with a situation in which there are exact zeros or ones would be to slightly perturb the data. In our case, that would likely have been defendable as there is not much difference between getting 99 per cent on a paper and getting 100 per cent. When we estimated that perturbed data model using beta regression the results were similar.

4 Results

Our central results are summarized in Table 6 and Figure 6, which use the `modelsummary` (Arel-Bundock 2022) package. These focus on the main estimates i.e. those respondents with less than full marks. The full output is available in Appendix B.1.

Table 6: Selected coefficient estimates and mean absolute deviation (MAD)

	STA302	STA302: GPA	STA304	STA304: GPA	STA304: GPA & ESL
Intercept	0.86	−2.06	0.91	−0.87	−0.95
	(0.17)	(0.49)	(0.22)	(0.41)	(0.43)
Intercept (Phi)	1.74	1.94	1.77	2.42	2.68
	(0.28)	(1.01)	(0.37)	(0.79)	(0.87)
Intercept (ZOI)	−2.10	−15.61	−2.93	−2.49	−3.07
	(0.61)	(4.08)	(1.13)	(2.39)	(2.60)
LLM usage: Somewhat	0.34	0.35	−0.05	−0.01	0.01
	(0.25)	(0.20)	(0.30)	(0.25)	(0.24)
LLM usage: Extensive	0.69	0.64	0.39	0.36	0.30
	(0.21)	(0.20)	(0.24)	(0.19)	(0.19)
LLM usage: Somewhat (Phi)	−0.05	−0.19	−0.29	−0.70	−0.75
	(0.38)	(0.41)	(0.47)	(0.47)	(0.47)
LLM usage: Extensive (Phi)	0.55	−0.12	0.21	−0.17	−0.40
	(0.37)	(0.39)	(0.38)	(0.41)	(0.42)
GPA		1.00		0.58	0.61
		(0.17)		(0.13)	(0.13)
GPA (Phi)		0.18		−0.05	0.00
		(0.33)		(0.25)	(0.26)
English native: Yes					0.08
					(0.19)
English native: Yes (Phi)					−0.78
					(0.29)
Num.Obs.	92	92	131	131	131
R2	0.094	0.478	0.052	0.164	0.171
ELPD	14.3	39.2	35.8	42.6	44.2
ELPD s.e.	10.5	9.4	13.6	15.3	15.1
LOOIC	−28.6	−78.3	−71.6	−85.1	−88.3
LOOIC s.e.	20.9	18.8	27.2	30.5	30.1
WAIC	−29.3	−79.0	−73.0	−86.9	−92.6
RMSE	0.15	0.11	0.16	0.15	0.15

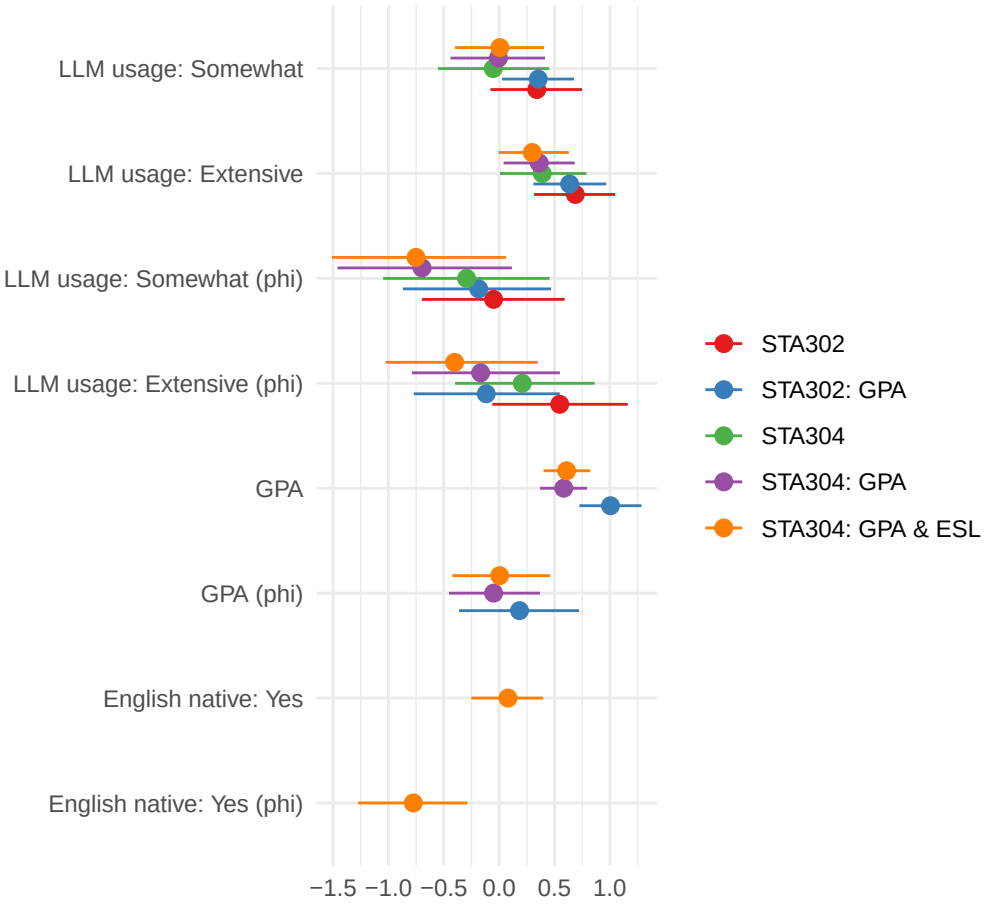


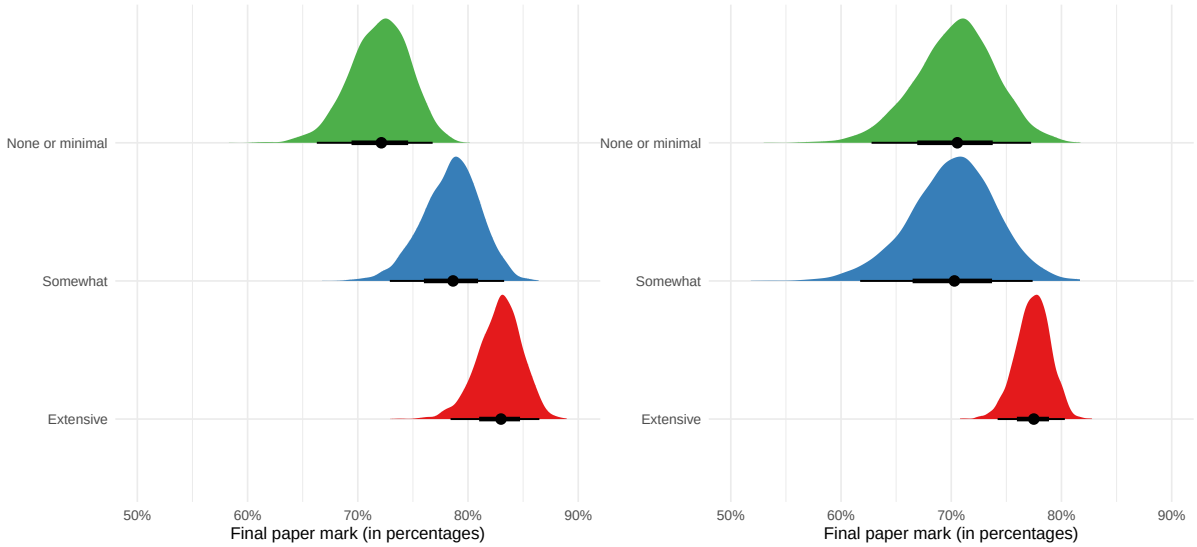
Figure 6: Selected coefficient estimates and 90 per cent credibility intervals

The initial model in Table 6 is for STA302 and focuses on the association between final paper mark and LLM usage. The next model adds in self-reported GPA. Both of these have 92 observations. This same pattern is then repeated for STA304 and both those models have 131 observations. The final model in Table 6 additionally considers whether the respondent is a native English speaker.

To begin with, we can take advantage of the simple initial models for STA302 and STA304 to look at the ZOI intercept. For instance, for the STA302 model it is -2.10. After undoing the logit transformation this is approximately 11 per cent, which corresponds to the number of students in STA302 who got full marks on the final paper (Heiss 2024). The same holds for the STA304 model.

LLM usage estimates are in relation to "None or minimal" LLM usage. In both the base STA302 and STA304 models, "Extensive" LLM usage was associated with slightly higher mean and precision estimates (Figure 6). But when GPA was added to the models the precision component became negative. That said, with 90 per cent credibility intervals there is considerable overlap with zero for many estimates.

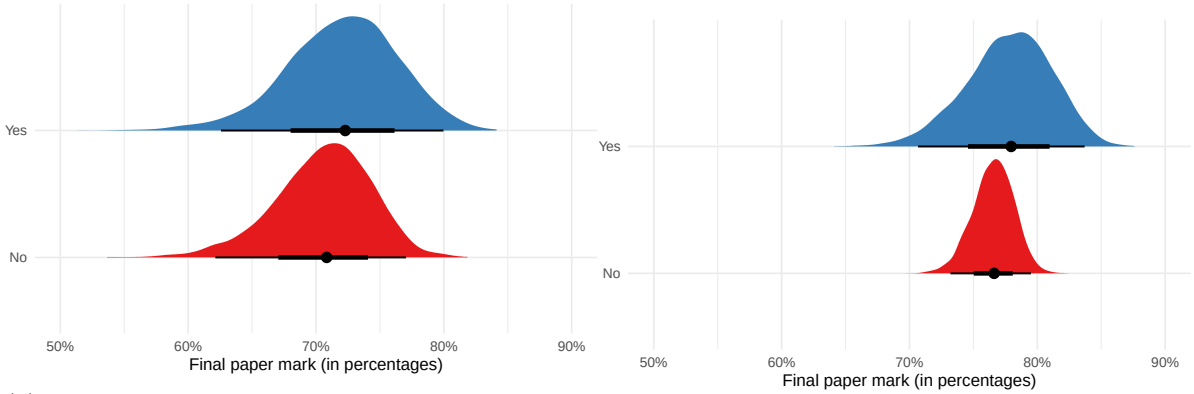
To understand the implication of the coefficient estimates we can hold all other variables constant and then vary LLM usage from "None or minimal" to "Somewhat" and then "Extensive" (Figure 7). There is an expected increase in the mark on the final paper as LLM usage increases in STA302 (Figure 7a), but it is less clear in STA304 (Figure 7b). It may have been that the overwhelming usage of LLMs at "Extensive" levels in STA304, which occurred six months after STA302, was the reason for this seemingly different association in our data.



(a) Based on responses from STA302 (Winter 2024) (b) Based on responses from STA304 (Fall 2024)

Figure 7: Comparing the effect of a hypothetical student with different LLM usage levels, all other variables constant, between STA302 and STA304

The different effect of LLM usage, by whether the student was a native English speaker, holding all other variables constant, can only be examined for STA304 (September-December 2024) as the question was not asked in the earlier STA302 (Figure 8). For the hypothetical student who does not speak English natively, where we changed nothing other than their LLM usage, going from "None or minimal" (Figure 8a) to "Extensive" (Figure 8b) appears to have been associated with higher marks in our dataset. The distribution is smaller and the average is higher.



(a) Holding GPA at 3.0 and LLM usage is none or minimal

(b) Holding GPA at 3.0 and LLM usage is extensive

Figure 8: Comparing the effect of a hypothetical student changing whether they were a native English speaker, all other variables constant, for STA304 only

5 Discussion

There may be some 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 further development, testing and integration in the data science process.
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 for students learning data science, particularly for those who are not native English speakers.

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. For instance, looking at the logs, a common use case was to ask the LLM to write code to make a graph, and another was to ask for some example paragraphs of text.

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 *prima facie* neither good nor bad in and of itself, however it can clash with the style of the code that surrounds it which could lead to maintainability difficulties.

Again, looking at the logs, 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 Stack Overflow 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 outdated 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 occurred 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 would likely require financial support.

Regression provides 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, considering STA302, 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. We were only able to distinguish between native and second-language English speakers at a high level and a more detailed focus on differential LLM usage would add a great deal of nuance to the analysis.

Finally, we only considered one outcome measure, namely a student's mark on their 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 workflow, both in and outside of the classroom.

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?

In STA304, students were additionally asked about whether they were an international student "Are you an international student?", whether English was their native language "Is English your native language (i.e. the language that you first spoke)?", and an attention check was included "Please select the option"Blue" ".

B Model details

B.1 Full model output

Table 7: Coefficient estimates and mean absolute deviation (MAD)

	STA302	STA302: GPA	STA304	STA304: GPA	STA304: GPA & ESL
b_Intercept	0.86	−2.06	0.91	−0.87	−0.95
	(0.17)	(0.49)	(0.22)	(0.41)	(0.43)
b_phi_Intercept	1.74	1.94	1.77	2.42	2.68
	(0.28)	(1.01)	(0.37)	(0.79)	(0.87)
b_zoi_Intercept	−2.10	−15.61	−2.93	−2.49	−3.07
	(0.61)	(4.08)	(1.13)	(2.39)	(2.60)
b_coi_Intercept	2.64	2.61	2.27	2.28	2.24
	(1.20)	(1.21)	(1.21)	(1.16)	(1.17)
b_llm_usageExtensive	0.69	0.64	0.39	0.36	0.30
	(0.21)	(0.20)	(0.24)	(0.19)	(0.19)
b_llm_usageSomewhat	0.34	0.35	−0.05	−0.01	0.01
	(0.25)	(0.20)	(0.30)	(0.25)	(0.24)
b_phi_llm_usageExtensive	0.55	−0.12	0.21	−0.17	−0.40
	(0.37)	(0.39)	(0.38)	(0.41)	(0.42)
b_phi_llm_usageSomewhat	−0.05	−0.19	−0.29	−0.70	−0.75
	(0.38)	(0.41)	(0.47)	(0.47)	(0.47)
b_zoi_llm_usageExtensive	0.25	0.09	−0.08	−0.13	0.12
	(0.79)	(0.93)	(1.22)	(1.30)	(1.31)
b_zoi_llm_usageSomewhat	−1.17	−1.48	−0.34	−0.43	−0.08
	(1.21)	(1.32)	(1.51)	(1.55)	(1.66)
b_what_is_your_gpa		1.00		0.58	0.61
		(0.17)		(0.13)	(0.13)
b_phi_what_is_your_gpa		0.18		−0.05	0.00
		(0.33)		(0.25)	(0.26)
b_zoi_what_is_your_gpa		3.99		−0.20	−0.21
		(1.14)		(0.69)	(0.74)
b_native_englishYes					0.08
					(0.19)
b_phi_native_englishYes					−0.78
					(0.29)
b_zoi_native_englishYes					1.01
					(0.87)
Num.Obs.	92	92	131	131	131
R2	0.094	0.478	0.052	0.164	0.171
ELPD	14.3	39.2	35.8	42.6	44.2
ELPD s.e.	10.5	9.4	13.6	15.3	15.1
LOOIC	−28.6	−78.3	−71.6	−85.1	−88.3
LOOIC s.e.	20.9	18.8	27.2	30.5	30.1
WAIC	−29.3	−79.0	−73.0	−86.9	−92.6
RMSE	0.15	0.11	0.16	0.15	0.15

B.2 Full coefficient estimates

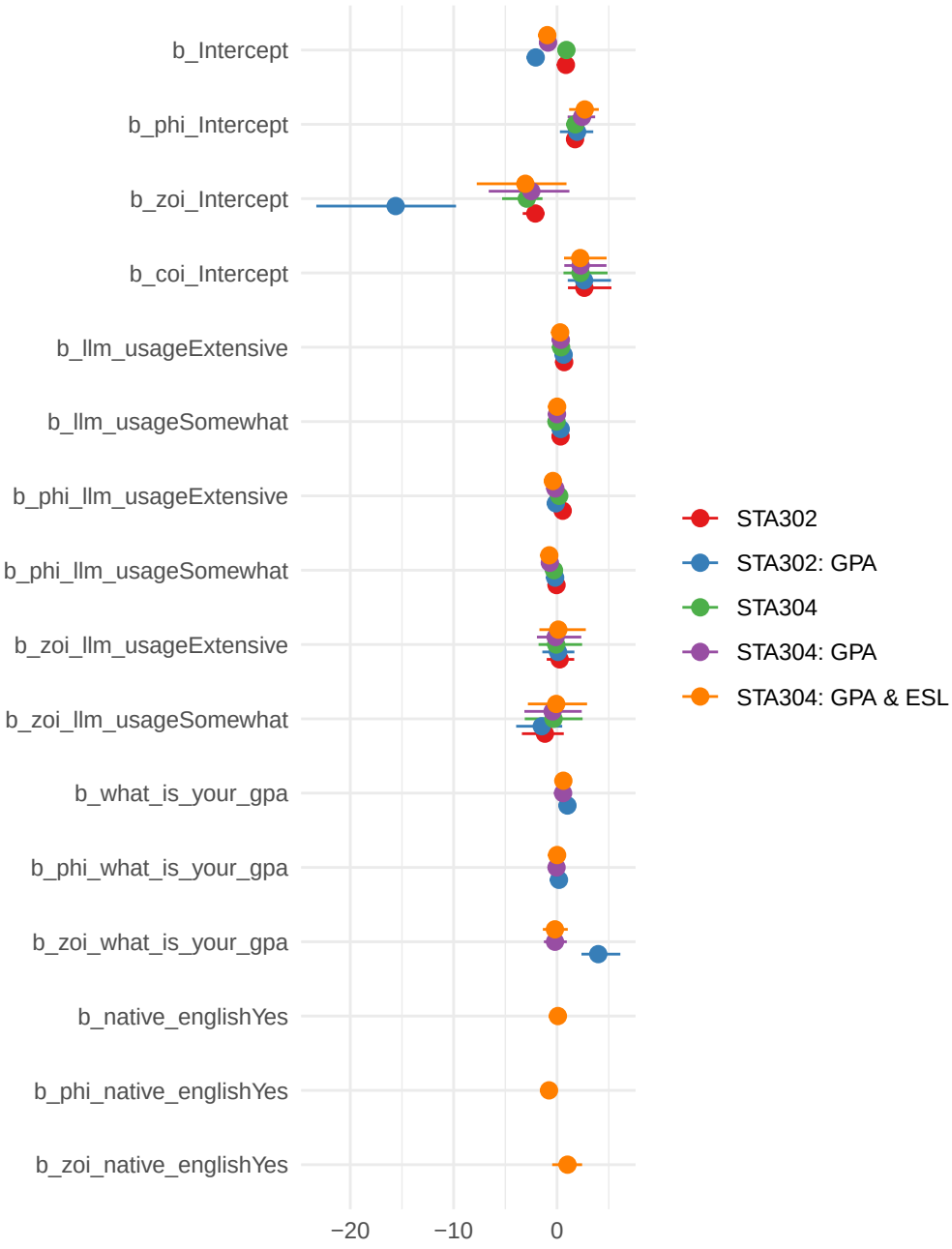
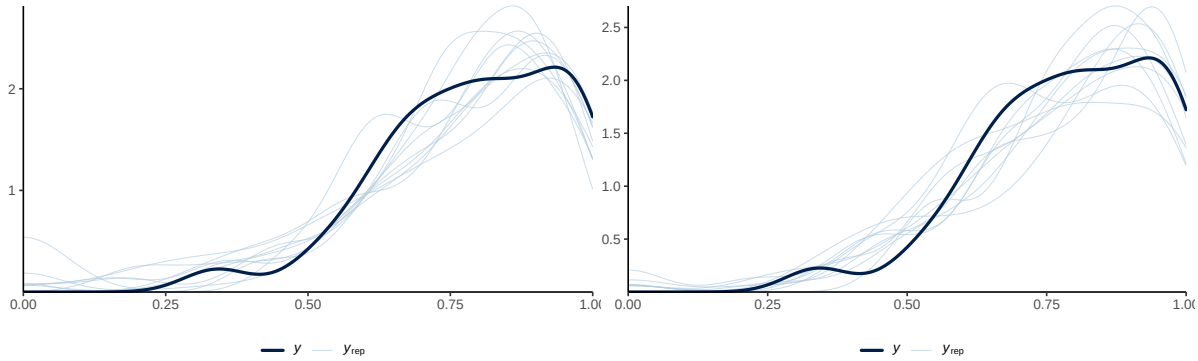


Figure 9: Coefficient estimates and 90 per cent credibility intervals

B.3 Posterior predictive check

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



(a) Base model (b) Including self-reported GPA

Figure 10: Posterior predictive checking

Table 8: Model comparison

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
fit2	0.000	0.000	39.171	9.382	11.686	2.021	-78.343	18.765
fit1	-24.854	6.547	14.318	10.457	8.765	1.603	-28.635	20.915

B.4 Diagnostics

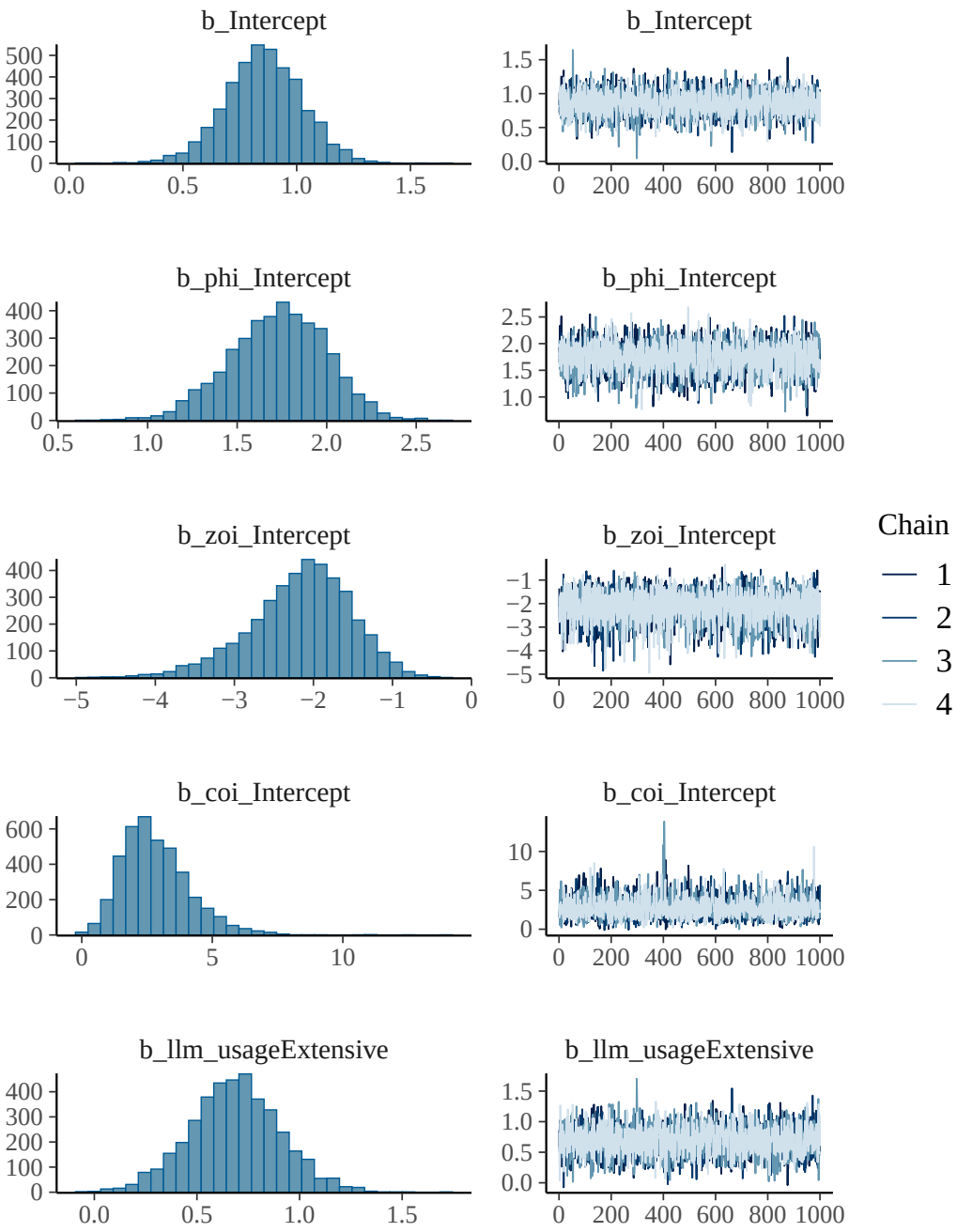


Figure 11: Base model diagnostics

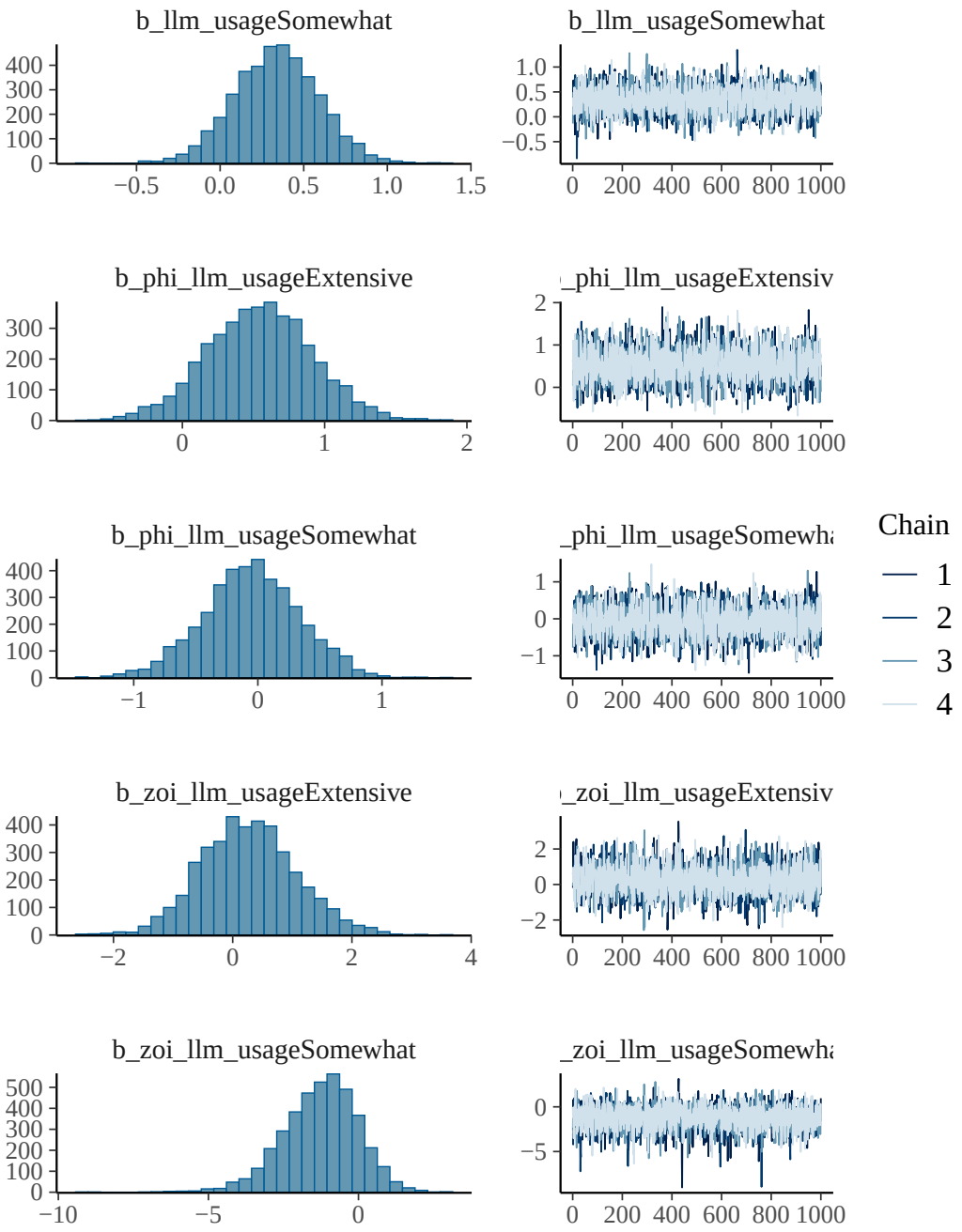


Figure 12: Base model diagnostics

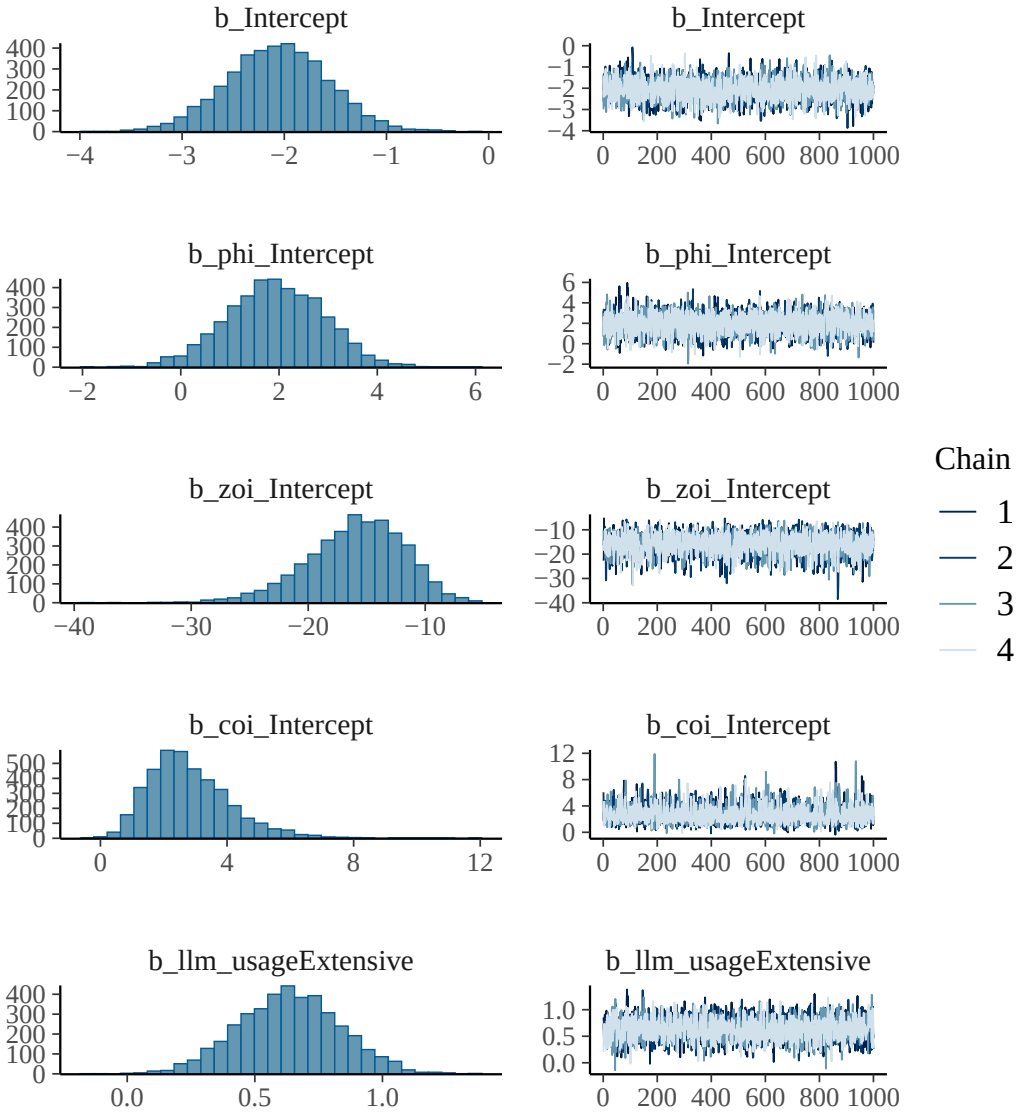


Figure 13: Model including self-reported GPA diagnostics

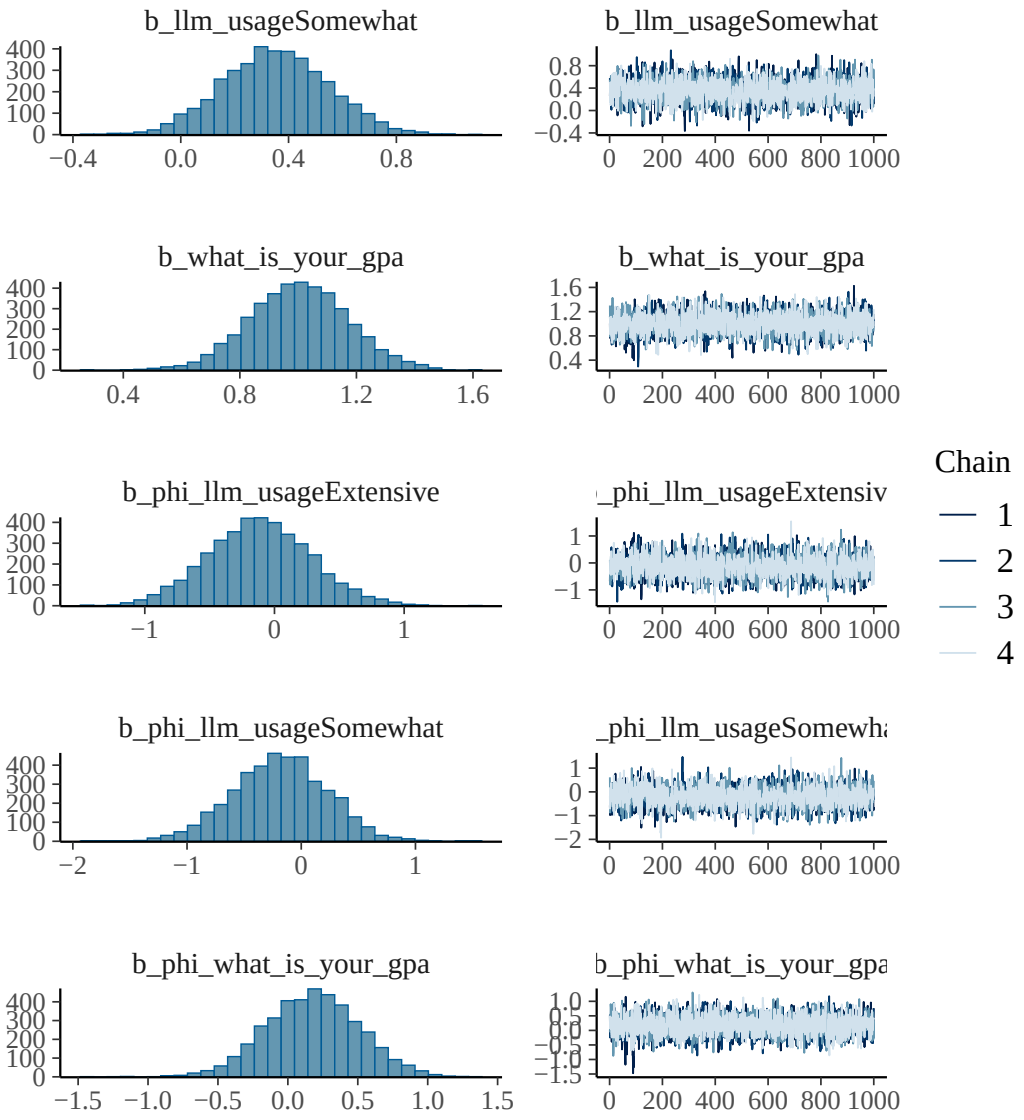


Figure 14: Model including self-reported GPA diagnostics

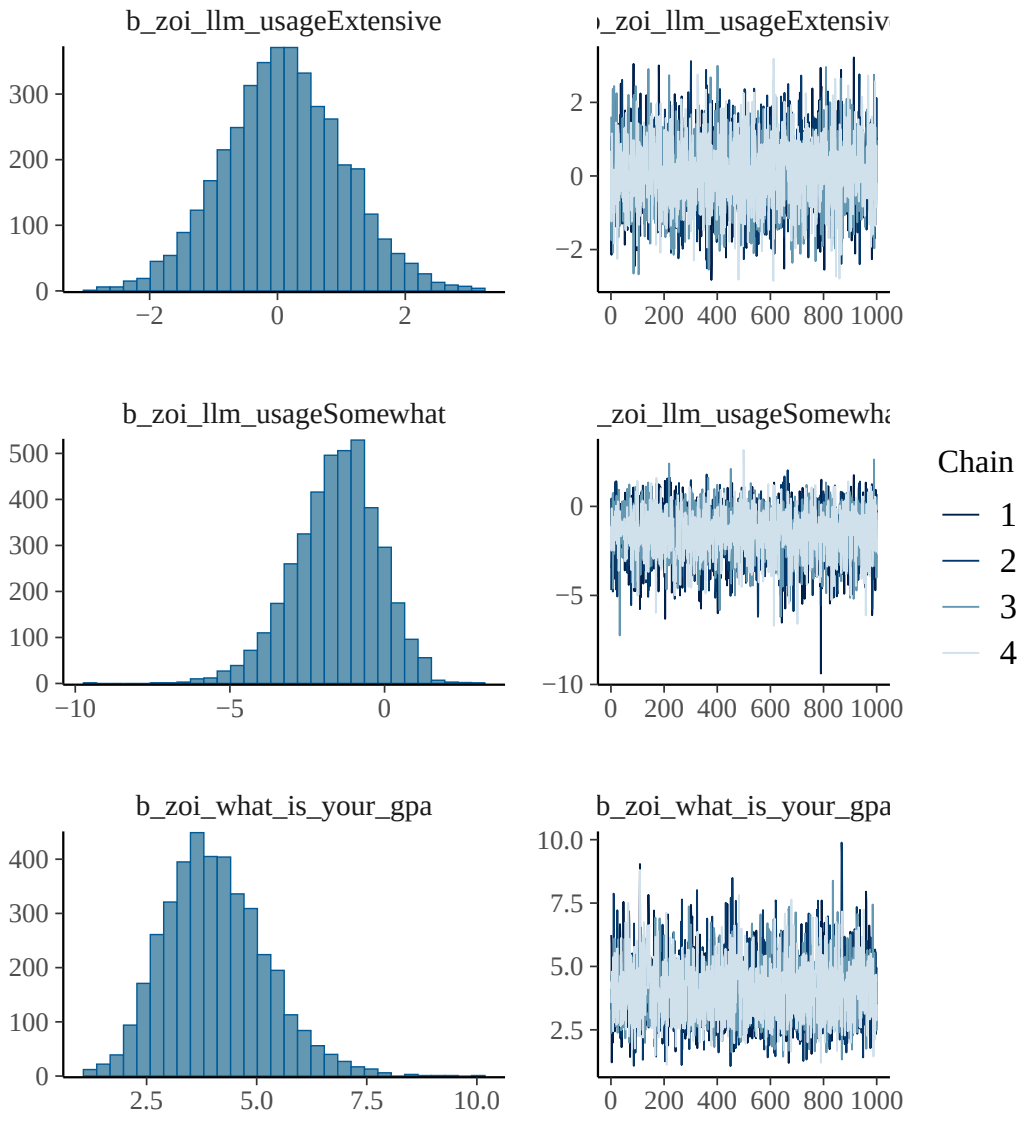


Figure 15: Model including self-reported GPA diagnostics

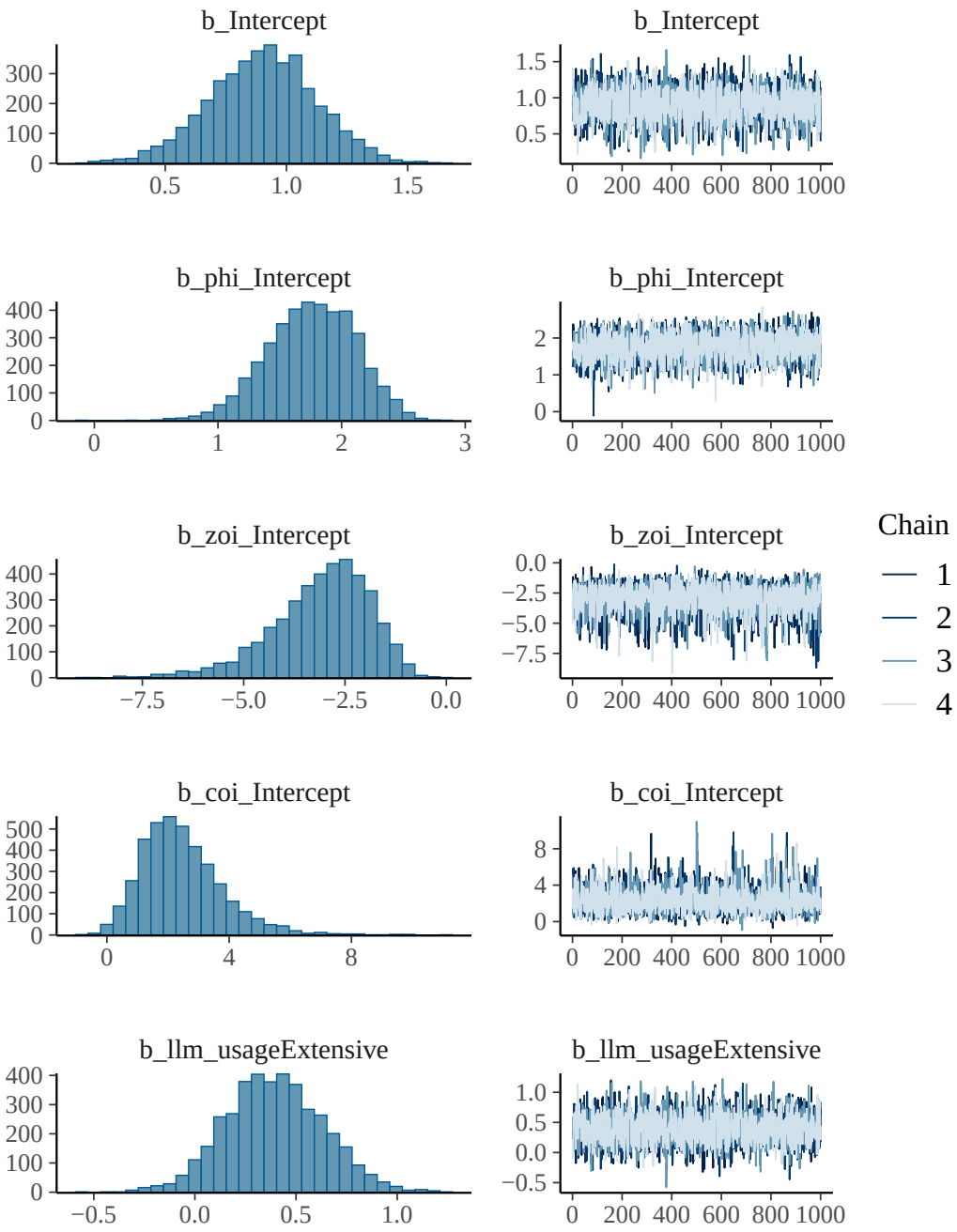


Figure 16: Base model diagnostics (STA304)

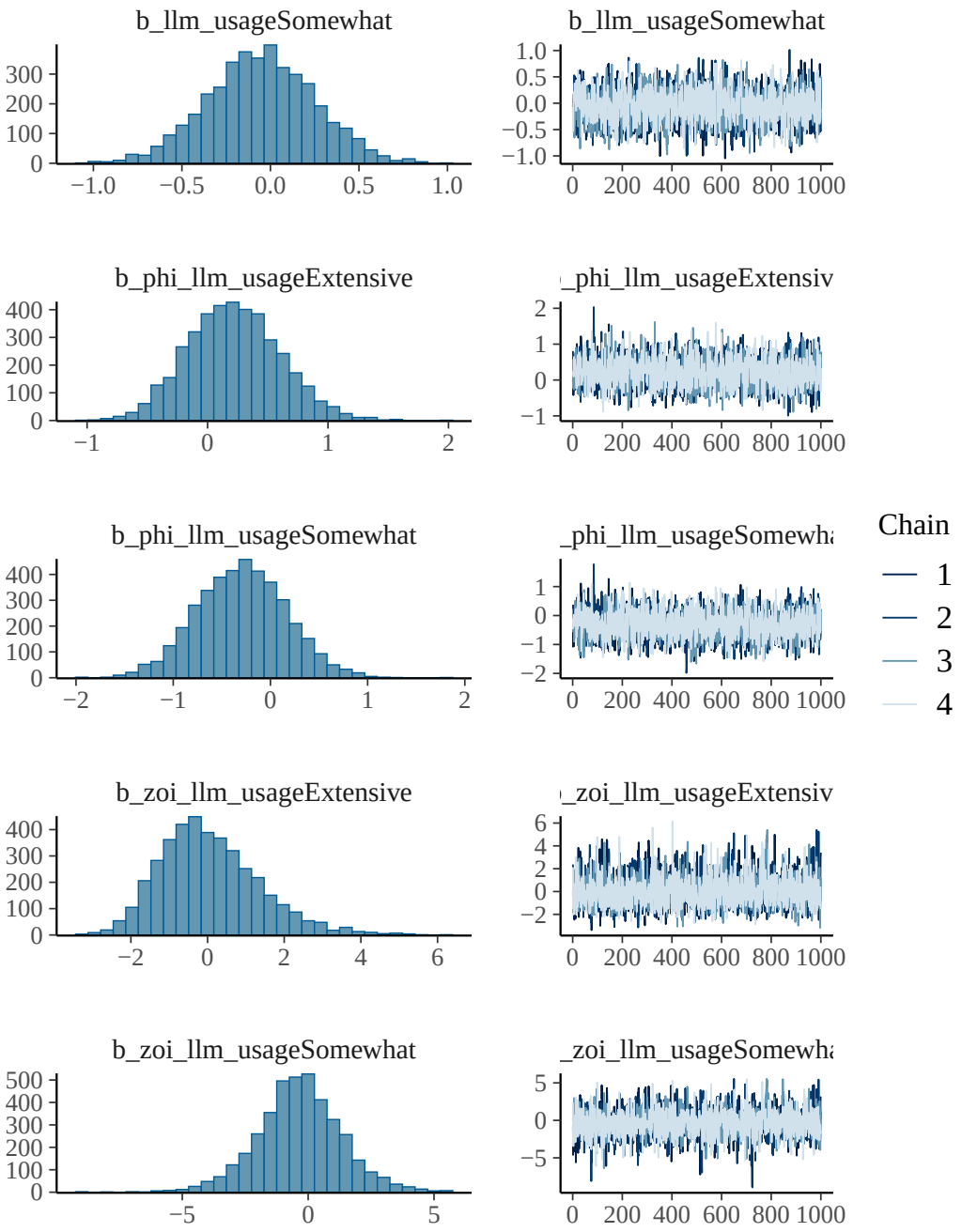


Figure 17: Base model diagnostics (STA304)

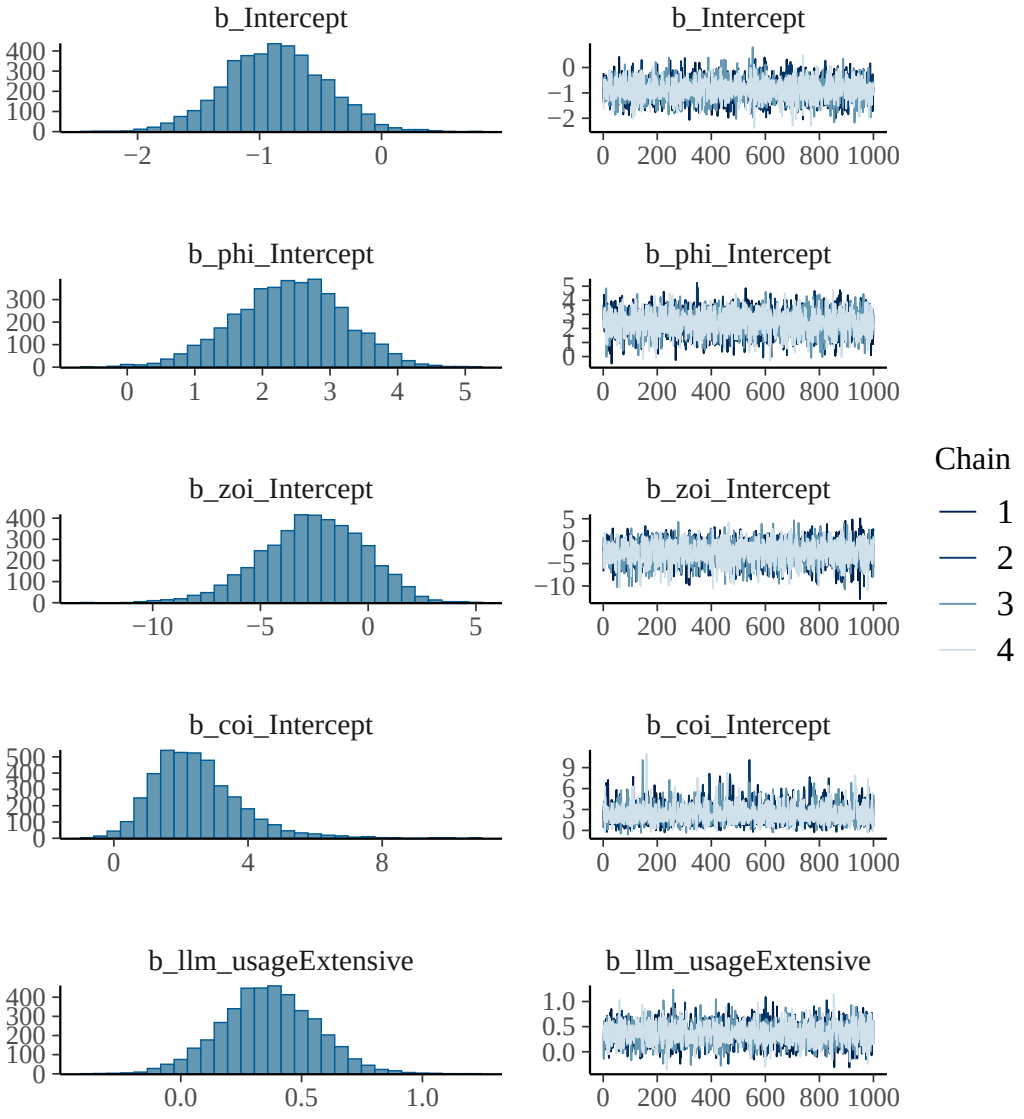


Figure 18: Model including self-reported GPA diagnostics (STA304)

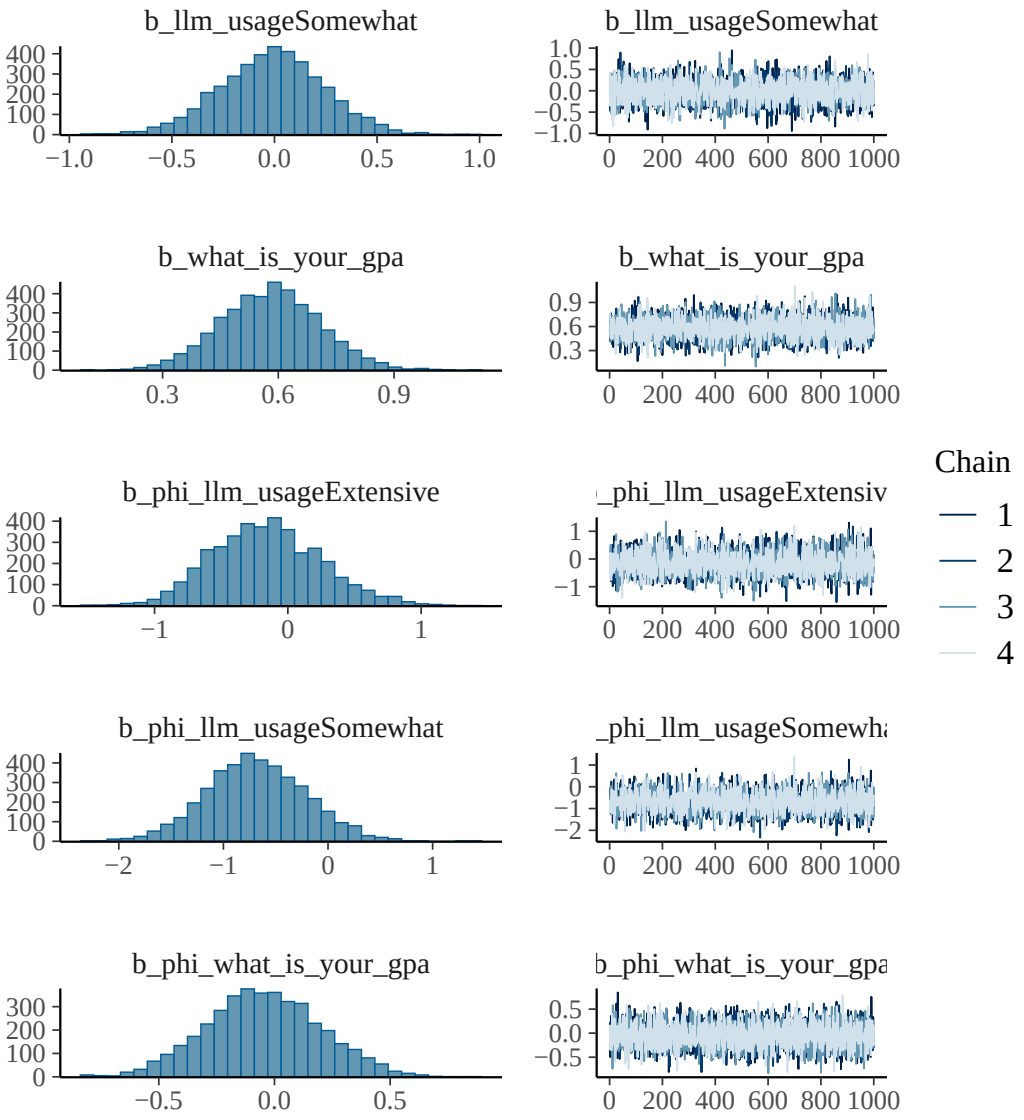


Figure 19: Model including self-reported GPA diagnostics (STA304)

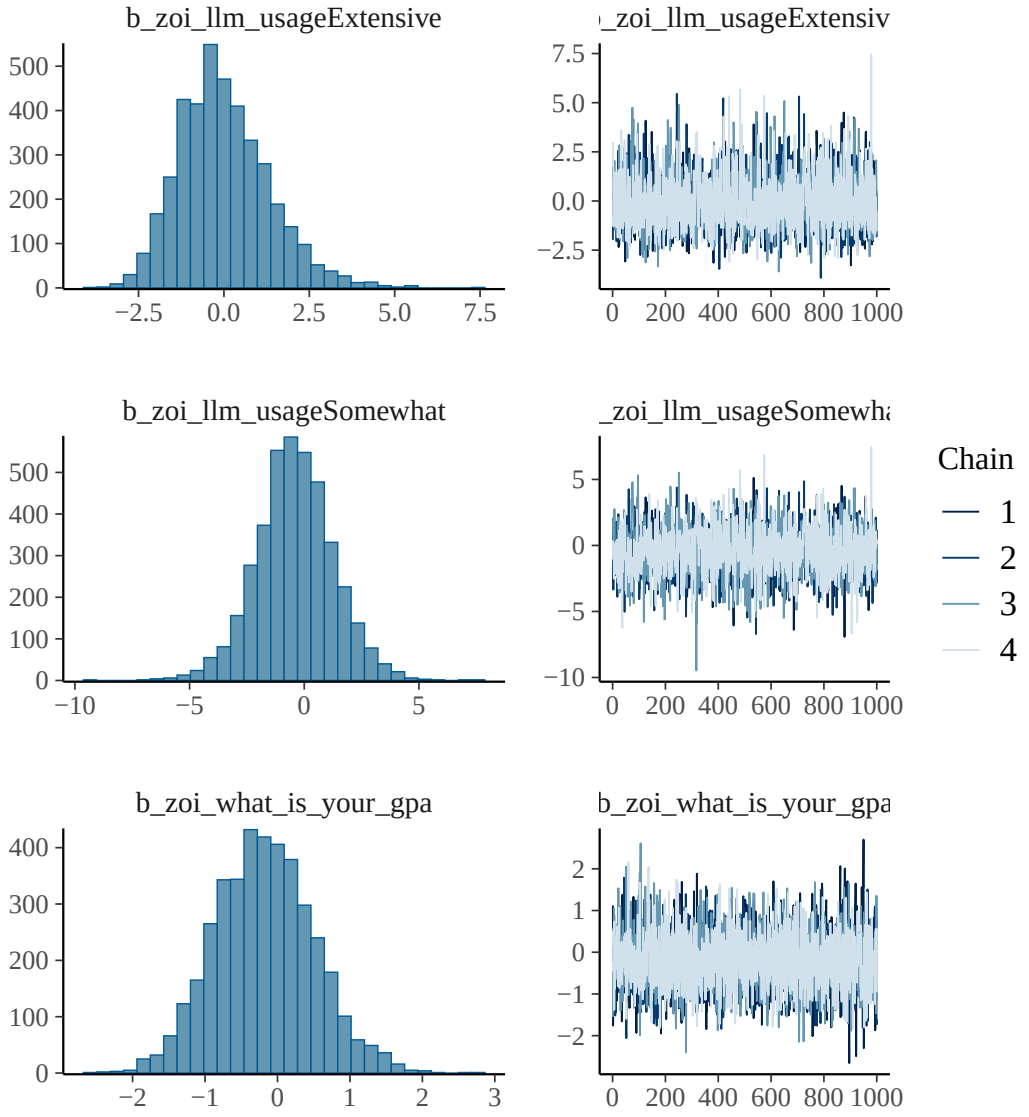


Figure 20: Model including self-reported GPA diagnostics (STA304)

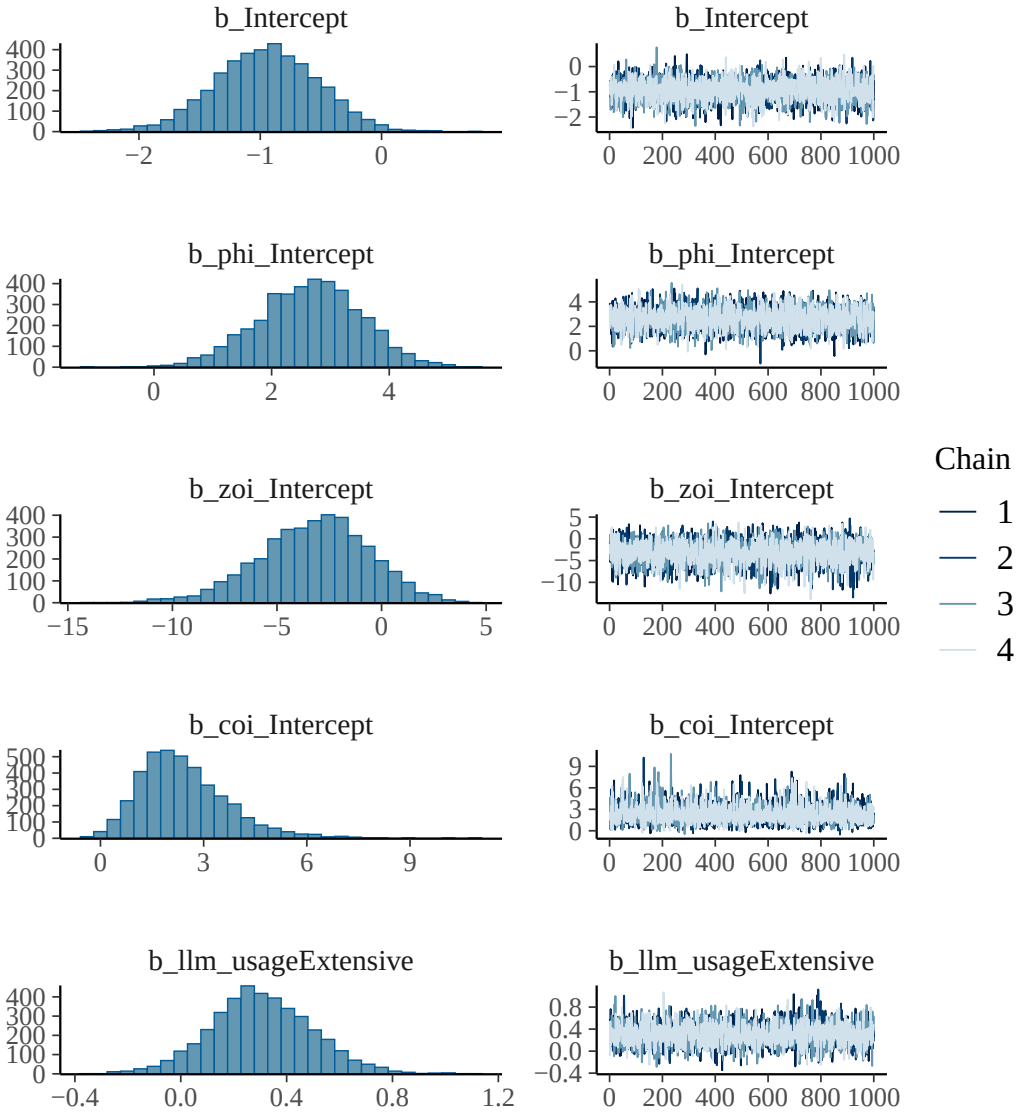


Figure 21: Model including self-reported GPA and ESL diagnostics (STA304)

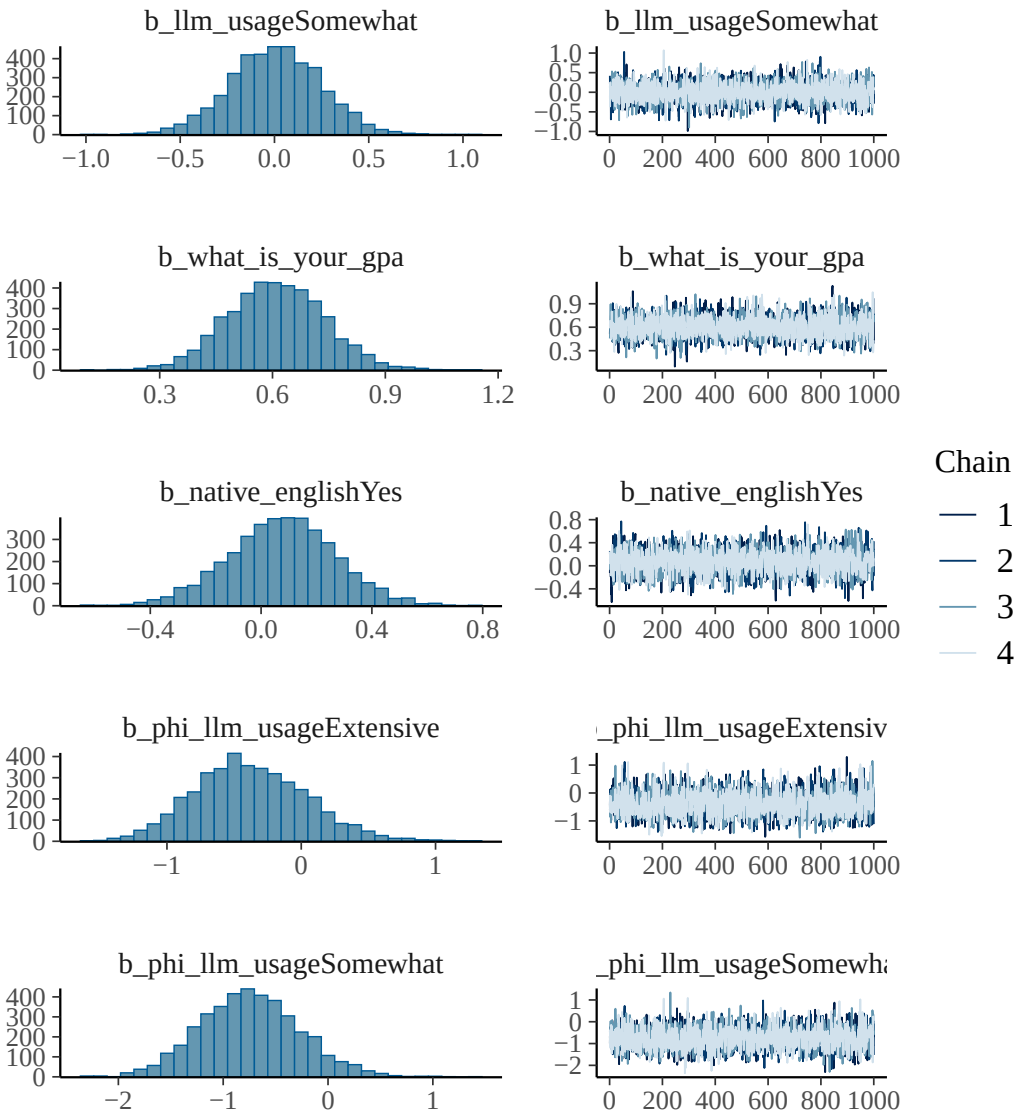


Figure 22: Model including self-reported GPA and ESL diagnostics (STA304)

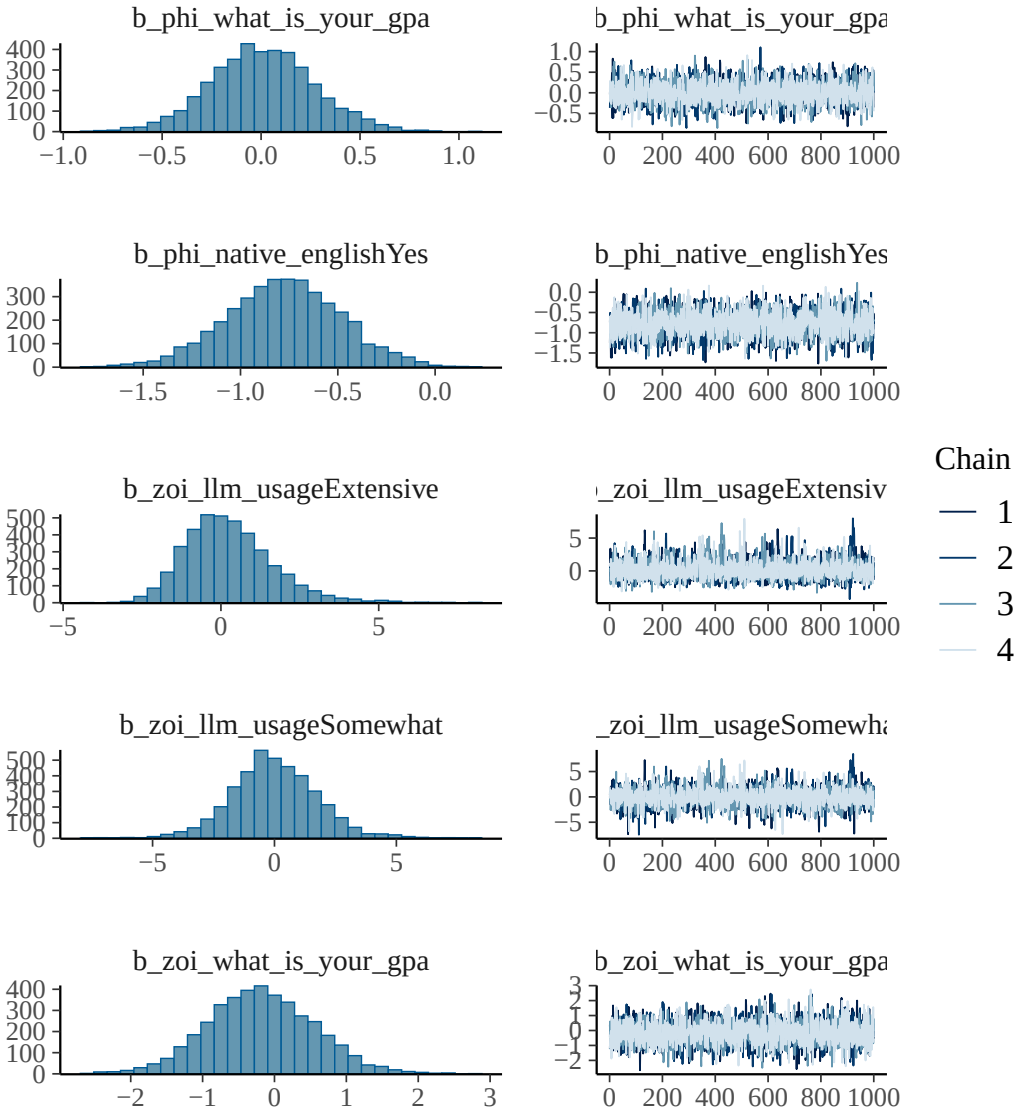


Figure 23: Model including self-reported GPA and ESL diagnostics (STA304)

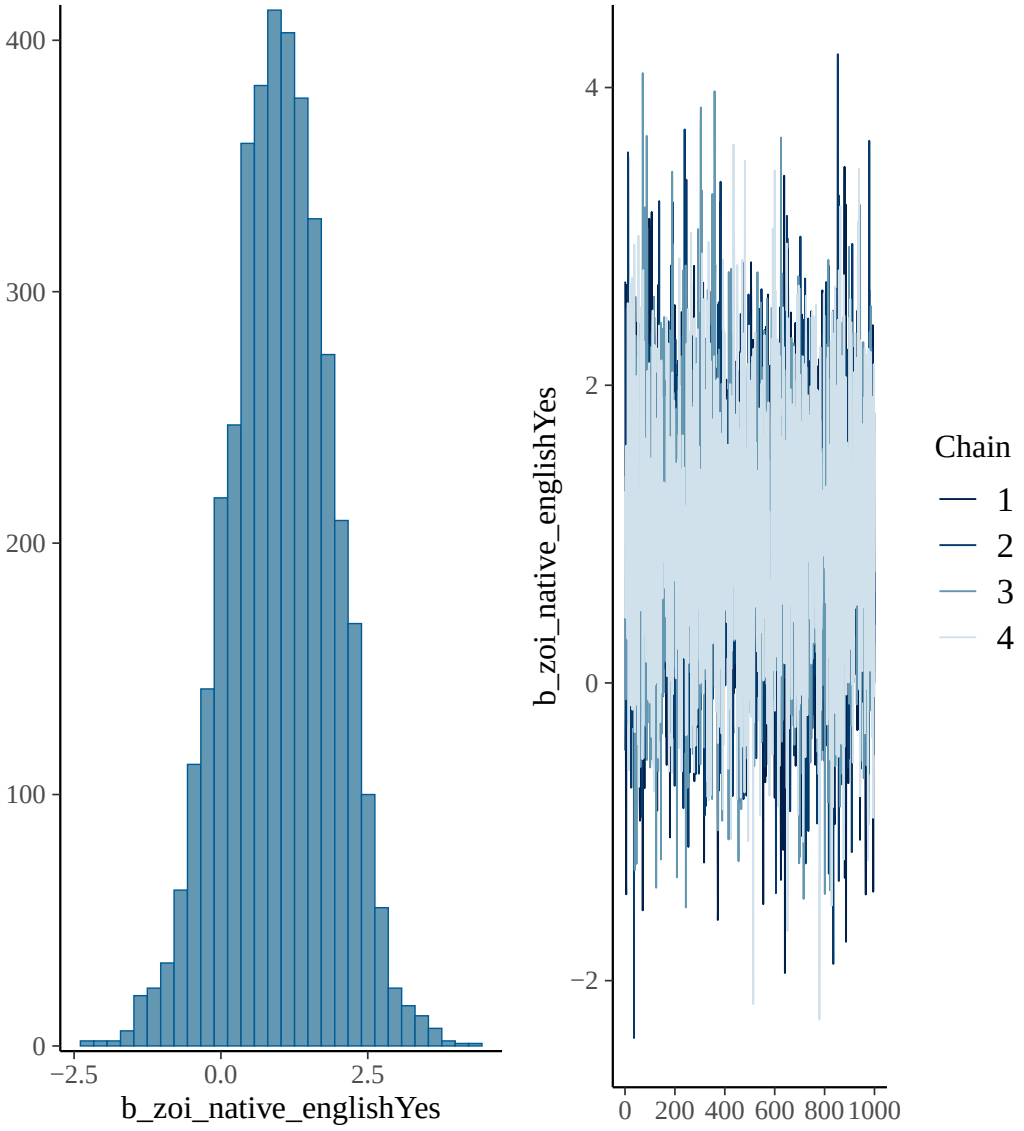


Figure 24: Model including self-reported GPA and ESL diagnostics (STA304)

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.

Anthropic. 2024. *I Want to Opt Out of My Prompts and Results Being Used for Training Models*. https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training.

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.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. "Generative AI Can Harm Learning." *The Wharton School Research Paper*, July. https://doi.org/10.2139/ssrn.4895486.

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.

———. 2020. "One-Inflated Beta Distribution," July. https://github.com/paul-buerkner/brms/issues/942.

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.

Heiss, Andrew. 2021. "A Guide to Modeling Proportions with Bayesian Beta and Zero-Inflated Beta Regression Models," November. https://doi.org/10.59350/7p1a4-0tw75.

———. 2024. "Zero/One-Inflated Beta Regression," April. https://talks.andrewheiss.com/2024-04-25_ksu-bayes/examples/zoib.html.

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.

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.

Lehmann, Matthias, Philipp B. Cornelius, and Fabian J. Sting. 2025. "AI Meets the Classroom: When Do Large Language Models Harm Learning?" https://arxiv.org/abs/2409.09047.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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.

Ospina, Raydonal, and Silvia L. P. Ferrari. 2012. "A General Class of Zero-or-One Inflated Beta Regression Models." *Computational Statistics &Amp; Data Analysis* 56 (6): 1609–23. https://doi.org/10.1016/j.csda.2011.10.005.

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. 2025. *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.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to "Apache" "Arrow"*. https://github.com/apache/arrow/.

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, 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.