TRUSTWORTHY DATA SCIENCE AND SELF-REPORTED LLM USAGE

EVIDENCE FROM A CANADIAN UNDERGRADUATE DATA SCIENCE COURSE

ROHAN ALEXANDER, 4 AUGUST 2024, JSM

(JOINT WORK WITH LUCA CARNEGIE)

Agenda

1. Motivation

2. Background

3. Data

4. Model

5. Results

6. Discussion

Main takeaway

1. Some very mild evidence from this observational study that LLM usage may be associated with better scores.

2. General purpose chatbot based LLM interactions mostly used for coding and writing, but not aspects that would enhance trust-worthiness.

3. Opinionated chatbots needed, as well as LLM based infrastructure.

Motivation

What makes **trustworthy** data science?

- Education

- Culture

- Workflow

- Tools:

  - Computers

  - Programming languages

  - Environments

November 30, 2022: A new tool!

- OpenAI's ChatGPT public release made LLMs mainstream. It is the chat aspect that was different.

- A rare instance of a new tool.

- LLMs can producing code and text, so they should be useful for trustworthy data science.

Evidence from a Canadian undergraduate data science course

- In this paper we are interested in better understanding LLMs as a tool for producing trustworthy data science.

- We study how LLMs were used by students in an upper-year undergraduate data science course.

- Do students who used LLMs have higher scores than those who did not?

Background

STA302 "Methods of Data Analysis I"

- Winter 2024 semester, University of Toronto.

- 275 students initially; 154 by the final paper.

- Assessment heavily based on three papers submitted over the course of the 12 week semester.

- Two previous papers have similar requirements and rubrics to that of the final paper.

- Students also have an, optional, two-day period of peer review.

Classroom ~= real world

- Papers reflect the work done by professional data scientists.

- A typical paper submission is 10-20 pages, and requires students to conduct original research to answer a research question of interest to them.

- Students:

  - develop a research question of interest to them,

  - identify or collect data to answer the question,

  - conduct statistical analysis, and

  - write a short paper.

Classroom ~= real world (evidence)

Hi Rohan, I would like to let you know that thanks to your final paper post, I was reached out by the Bank of Canada a while back, and now I will be working with them! They were impressed that we used Bayesian inference in our STA302 final papers and the emphasis on reproducibility and the data science workflow.

I was hoping to find an opportunity to tell you that I'm so happy I took STA302 with you because I got both of these jobs due to the course. Those long nights writing papers really paid off! My manager read my Toronto homicide paper and found it interesting, which helped me get the interview. My manager looked into my GitHub portfolio and hired me because he wanted someone with data analysis skills on his team (even though most of my tasks so far have been in Excel 😂).

Example paper

Fake News vs Fox News*

The Influence of Media Preferences on Voting Behavior in the 2020 U.S. Presidential Election Among Party Voters

Hannah Yu

April 19, 2024

Media consumption has a notable impact on voting behaviour, with the most influential example in the 2020 US presidential elections. Using logistic regression and data from the Cooperative Election Study (CES 2020), I analyze the relationship between watching specific media networks and voting for either Trump or Biden among Democrat, Republican, Independent, and Other voters. The analysis shows that viewing political-leaning networks significantly impacts voting preferences among voters from all party backgrounds, with CNN viewers more likely to support Biden and Fox News viewers favouring Trump. The findings highlight the media's role in shaping electoral outcomes, especially its influence on Independent voters, and the importance of addressing challenges posed by media polarization.

Example GitHub repository

Fox-News Public   👁 Watch 1 ▾

🔀 main ▾   ⑂ 1 Branch   🏷 0 Tags   🔍 Go to file   t   Add file ▾   <> Code ▾

👤 hannahyu07 change graph title   b280815 · 4 months ago   🕐 29 Commits

📁 data   more fixes, add graphs.   4 months ago
📁 models   more fixes, add graphs.   4 months ago
📁 other/llm   delete unwanted graphs   4 months ago
📁 paper   change graph title   4 months ago
📁 scripts   fix abstract   4 months ago
📄 .gitignore   set up   4 months ago
📄 Fox-News.Rproj set up   4 months ago
📄 README.md   Update README.md   4 months ago

📖 README   ✏ ☰

Fake News vs Fox News: The Influence of Media

Example LLMs statement

- `scripts` contains the R scripts used to simulate, download, test, and clean data.

Statement on LLM usage

Aspects of the code and paper were written with the help of ChatGPT. Some of the data interpretation, introduction, abstract and discussion were also written using ChatGPT. The entire chat history is available in `inputs/llms/usage.txt`

Students encouraged to use LLMs

- Formal instruction was provided twice during the semester:

  1. A masterclass taught by a computer science faculty member on the ethics of using LLMs.

  2. A masterclass taught by a TA on writing with LLMs.

- Students were required to disclose their usage through a statement in the GitHub repository README for the paper.
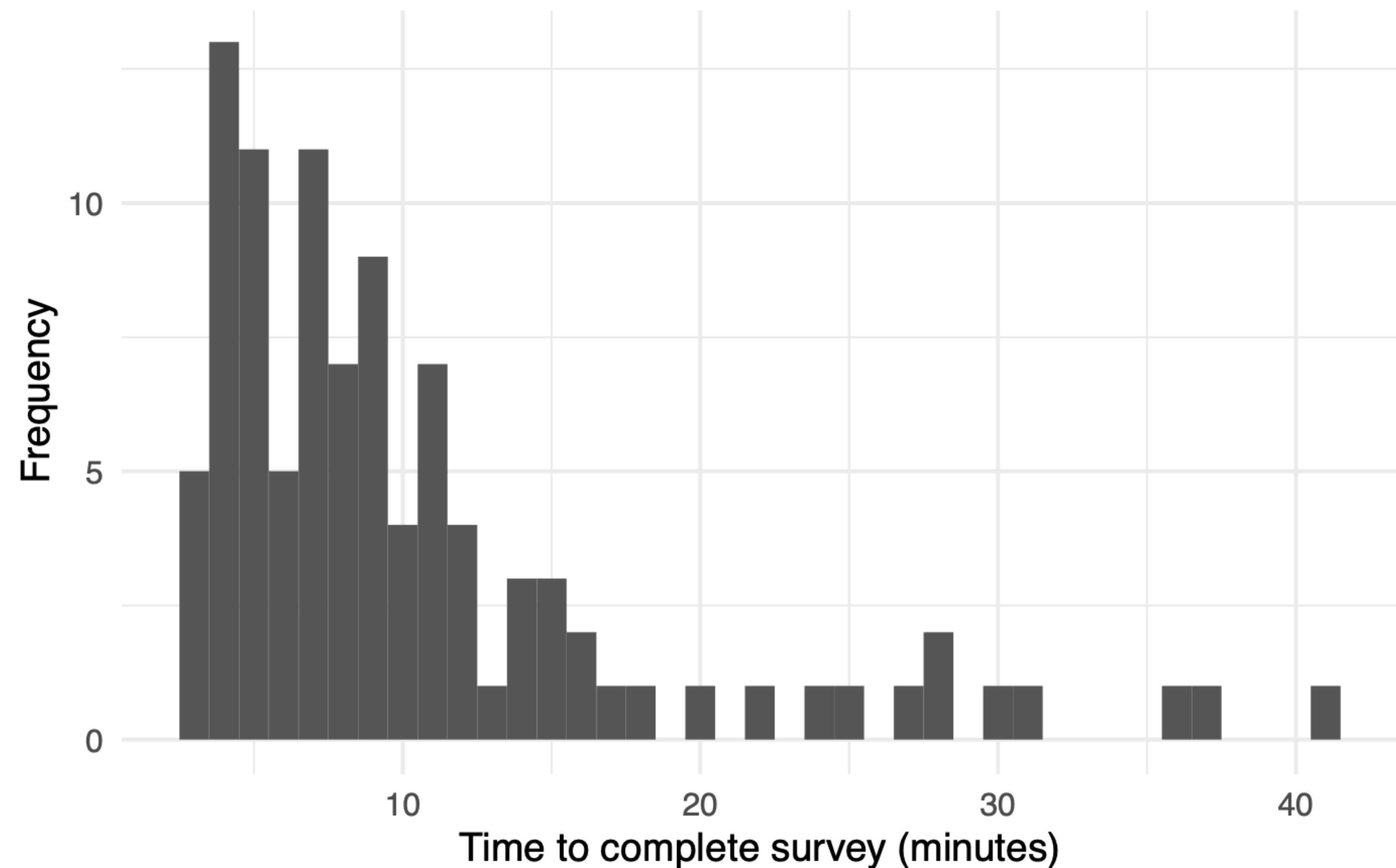
Data

Dataset: Overview

- Dataset containing student usage/attitudes, coursework, and academic performance was constructed.

- This was based on three components:

  1. An optional (but rewarded) survey.

  2. Self-reported LLM usage.

  3. Student marks on the final paper.
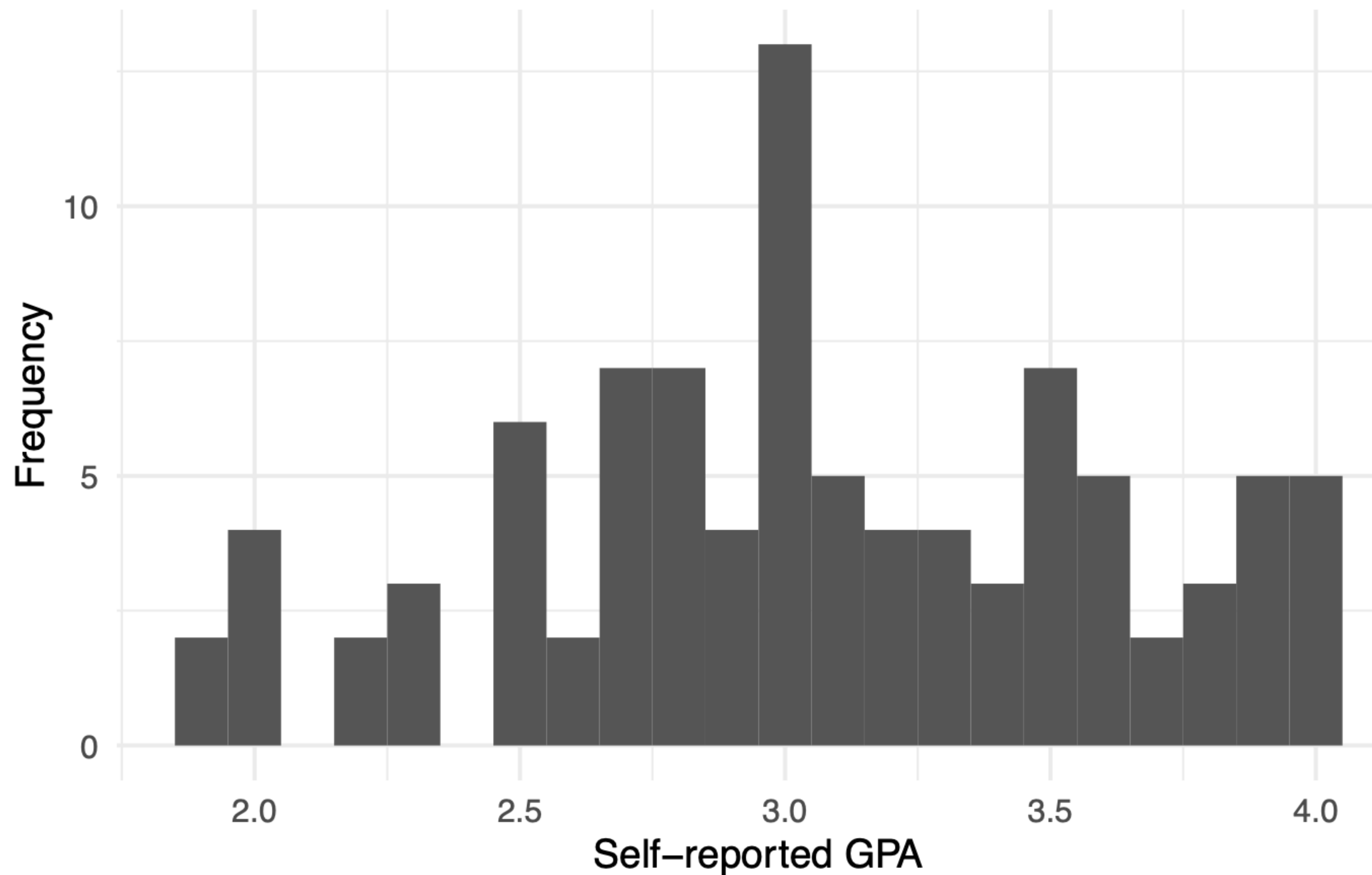
Dataset: Survey

(Thank you to Nathalie Moon for developing and sharing the questions.)

- 100 usable responses:

  - 146 responses to the survey.

  - 119 respondents provided authorization.

  - 4 respondents submitted the survey twice.

  - 15 of the remaining respondents did not include a usable LLM statement.

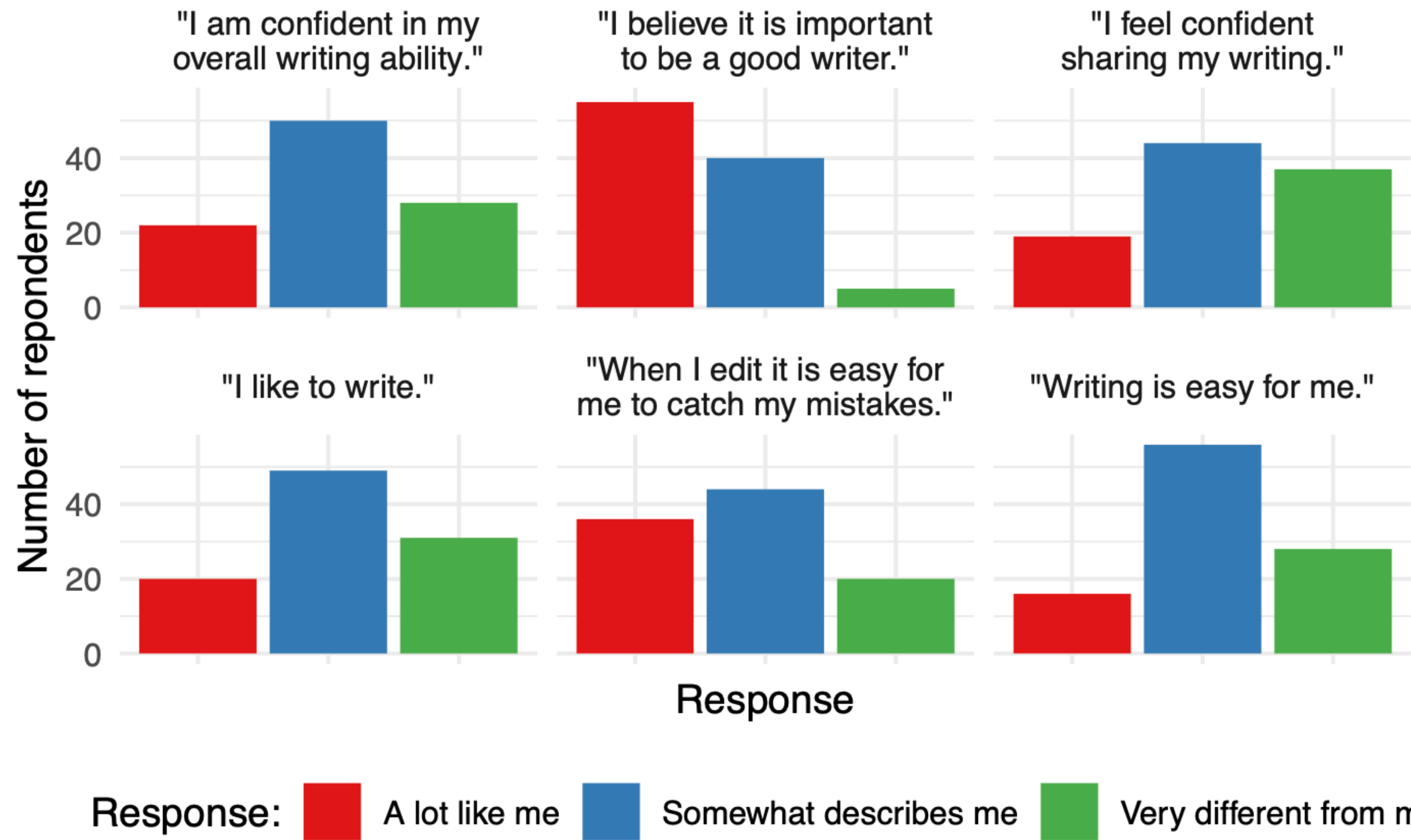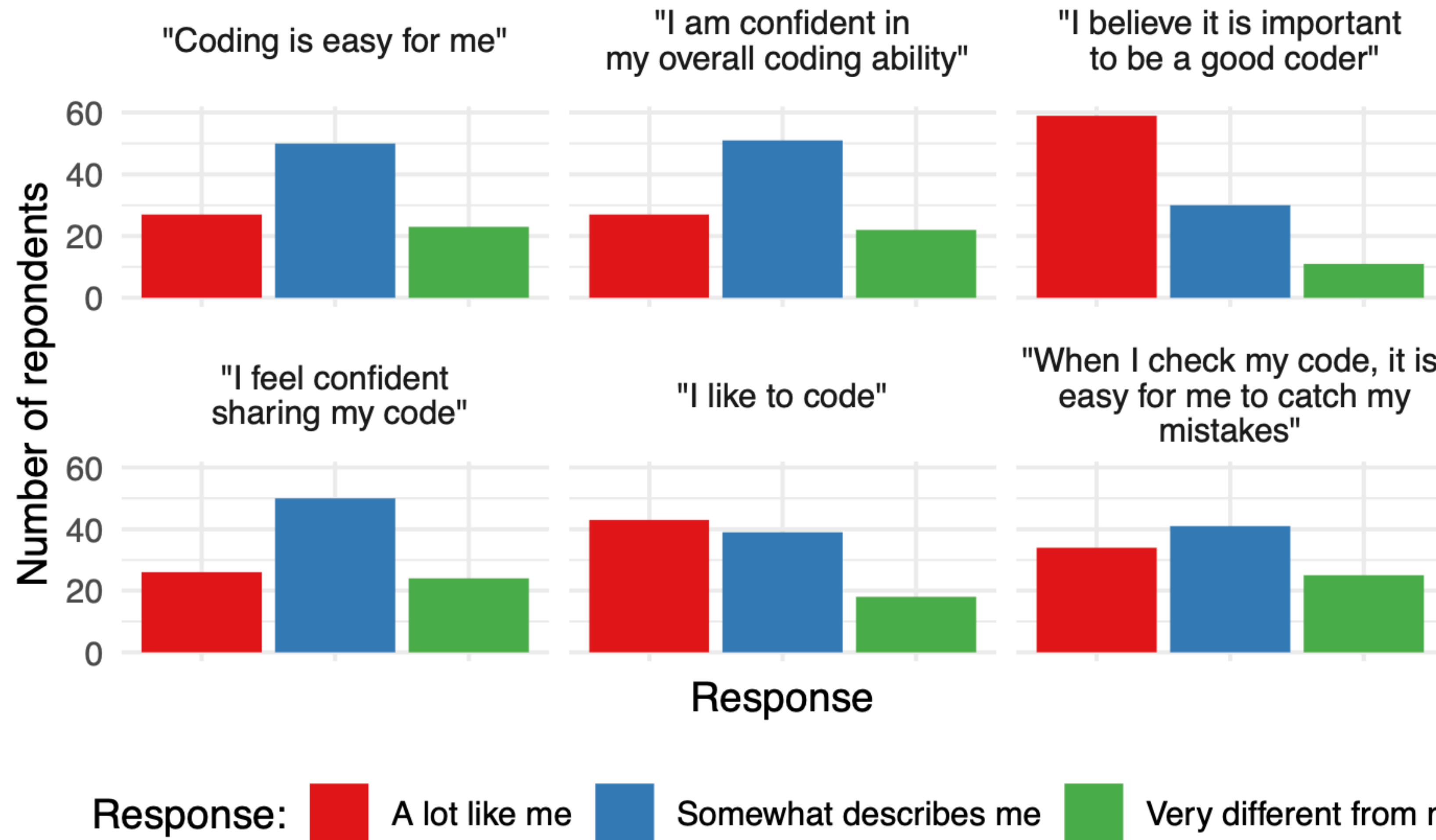- 2x 2nd years, 52x 3rd years, 42x 4th years, and 2x 5th year or over.

Dataset: Survey

Dataset: Survey

(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' "

Dataset: Survey

(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'"

Dataset: Survey

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

Extent of AI familiarity	Number
Not familiar	2
Somewhat familiar	56
Very familiar	42

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

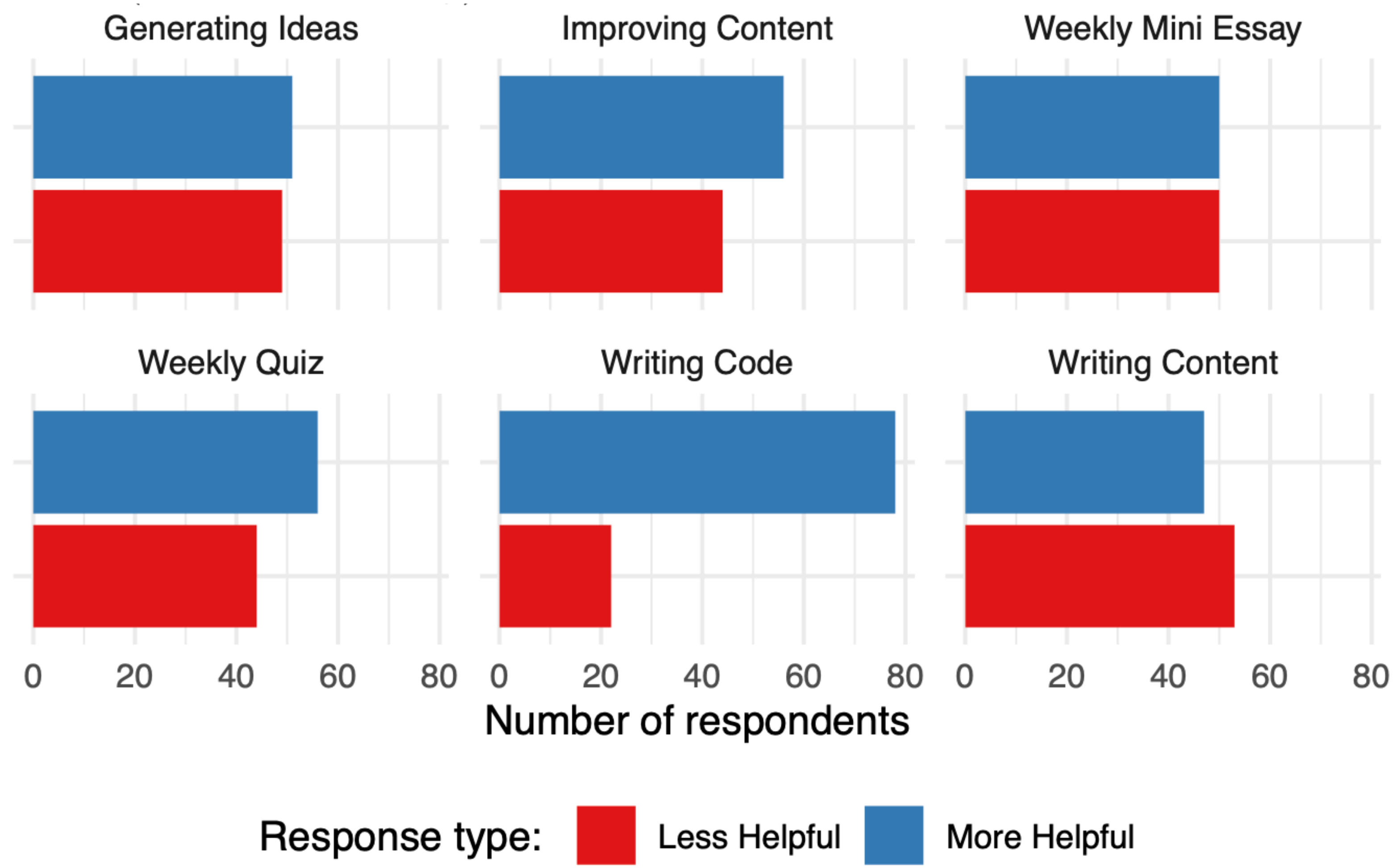| Ethical and appropriate for school?	Number
Appropriate	80
Inappropriate	9
It depends	11

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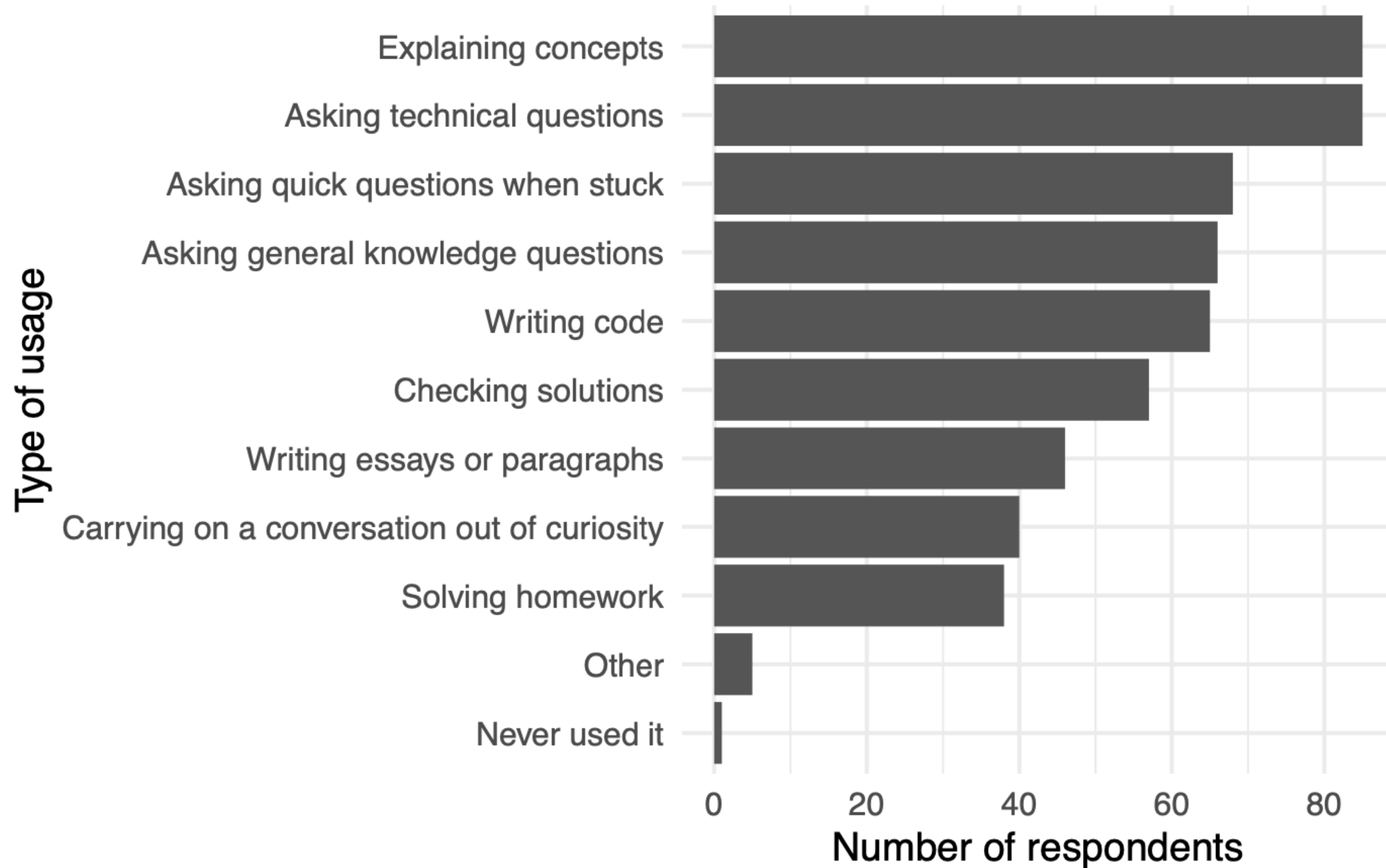
(a) "If you have used generative AI tools such as OpenAI's ChatGPT or equivalents, in what ways have you used it (select all that apply)?"

Dataset: Survey

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

Dataset: Self-reported LLM usage

- Even students who did not use generative AI at all were required to state this in the README.

- 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", "Some- what", "Extensive", "Unsure". Respond with only one of those options.'

Self-reported LLM usage	Number
Extensive	41
Minimal	8
None	23
Somewhat	28

Dataset: Final paper mark
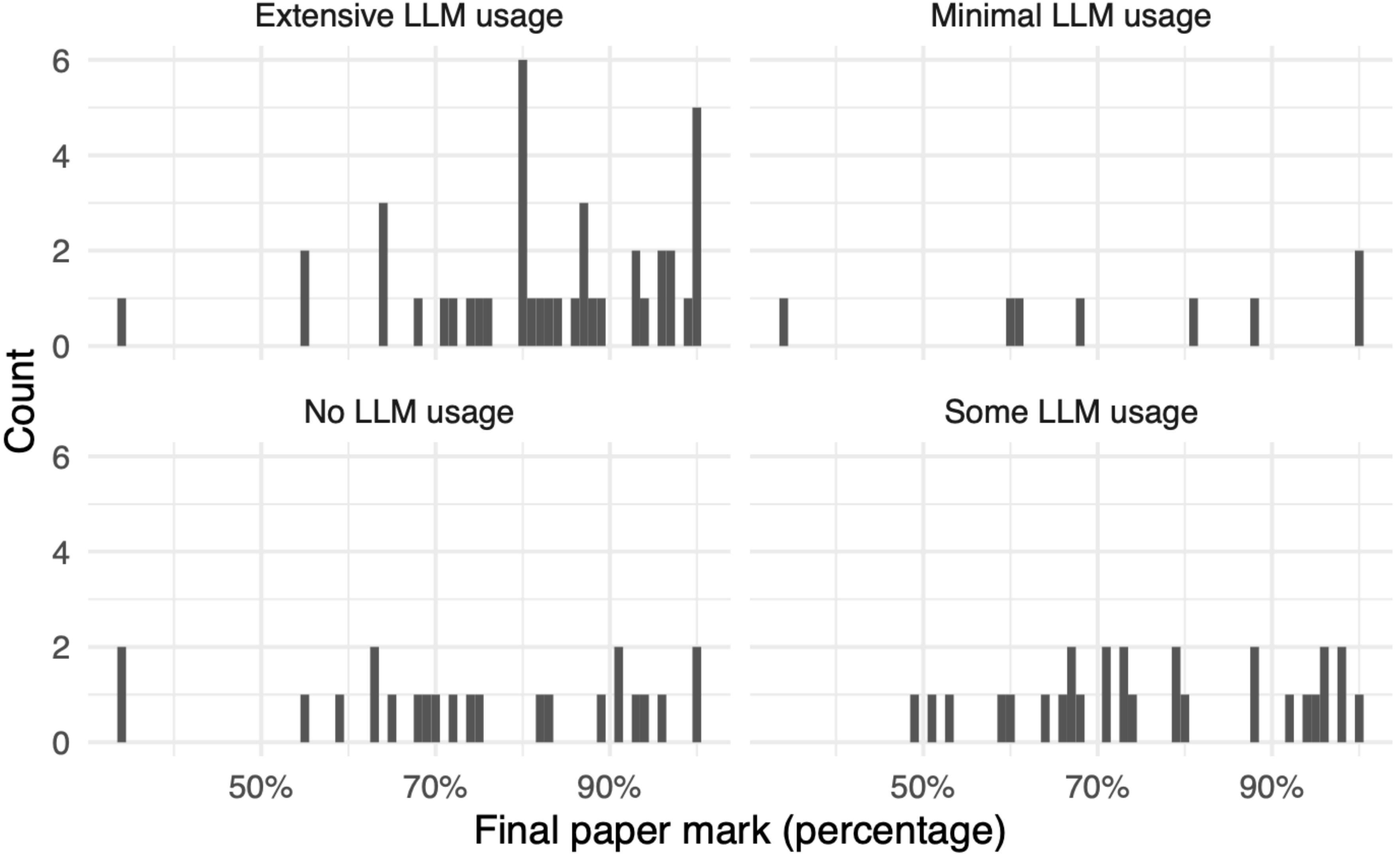
Model

Model: Zero-one-inflated beta regression

- The goal of our modelling strategy is to better understand how respondents result on their final paper is associated with their self-reported LLM usage.
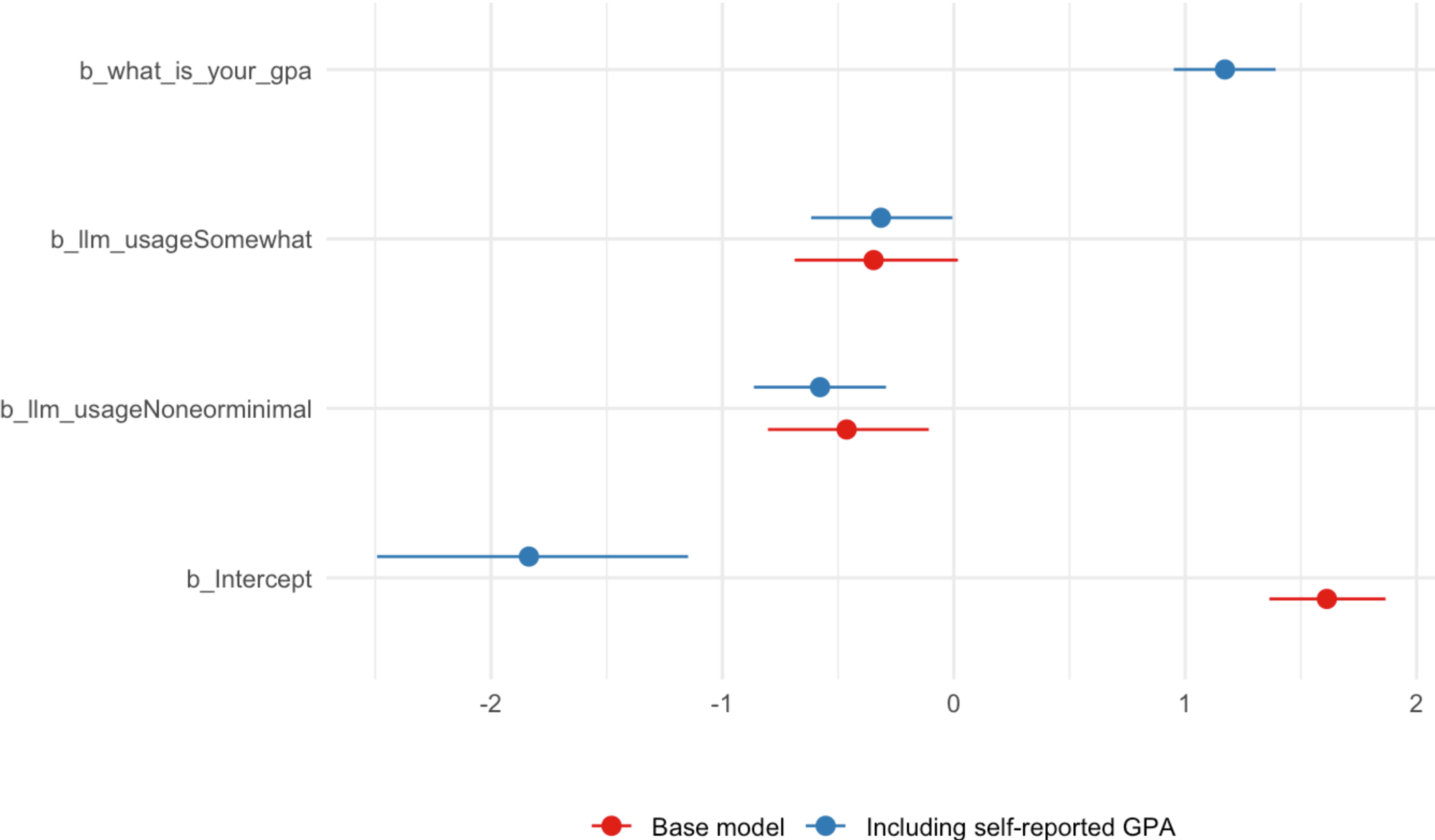
$$y_i \sim \text{Beta}(\mu_i, \phi) \tag{1}$$

$$logit(\mu_i) = \beta_0 + \beta_1 \times \text{LLM usage}_i + \beta_2 \times \text{GPA}_i \tag{2}$$

$$\beta_0, \beta_1, \beta_2 \sim \text{Normal}(0, 1) \tag{3}$$

$$\phi \sim \text{Gamma}(4, 0.1) \tag{4}$$

Results: Coefficient estimates and 90 % credibility intervals

Discussion

Discussion and weaknesses

- ChatGPT's chat interface made it popular, needs to be more opinionated and there needs to be more infrastructure developed.

- Low-resource languages, and data science style code, not well supported.

- Distributional effect: It may be that the results are different for different tranches of respondents. For instance, looking at the 28 students who received an A+ for the final paper, 13 of them had extensive LLM usage.

- However, this was not an RCT and much of the data were self-reported.

Thank you!

Rohan Alexander

rohan.alexander@utoronto.ca

rohanalexander.com