Using large language models to summarize student feedback

Presenter:
Nicholas T. Young[1]

Christopher Overton[1], Ania Majewska[2], Hina Shaikh[1,3], Nandana Weliweriya[1]

[1] Department of Physics and Astronomy, University of Georgia
[2] Department of Physiology and Pharmacology, College of Veterinary Medicine, University of Georgia
[3] Institute for Astronomy and Astrophysics, Eberhard Karls University of Tübingen

Introduction

- Collecting and acting on student feedback is an important method for instructors to adapt their teaching to student needs [1].
- Analyzing feedback from students in large-enrollment introductory courses, such as introductory physics at large universities, can be time-consuming for instructors.
- Generative AI tools are effective at producing summaries of text [2, 3] and therefore offer a potential solution for instructors to quickly extract key points from student feedback.
- Here, we compare generative AI's ability to extract key themes and trends from student feedback compared university instructors.

Methods

Data: Student responses to two open-response questions on the end-of-course evaluation-of-teaching survey from 8 introductory physics courses taught by 3 unique instructors.

1. *What do you feel are your instructor's strengths and weaknesses?*
2. *What do you feel are the strong and weak aspects of the course?*

Prompt [4]: "*For responses to open-ended questions, the goal is to focus on the useful information and identify trends or themes that appear. Note the frequency of themes, areas of agreement and disagreement among students, and suggestions students have for changes you might make. Please ignore the comments that are nonspecific. For the remaining comments, please sort them into three categories: positive, actionable suggestions, and nonactionable suggestions before identifying trends or themes*"

5 times each

5 Instructors

ChatGPT 4o

Claude 3.5 Sonnet

Gemini 1.5

Llama 3.1 70B

Identify themes noted by at least 3 of the 5 instructors

Identify themes noted by at least 3 of the 5 model runs for each of the four AI models.

Compare, taking themes identified by instructors as the "true" themes

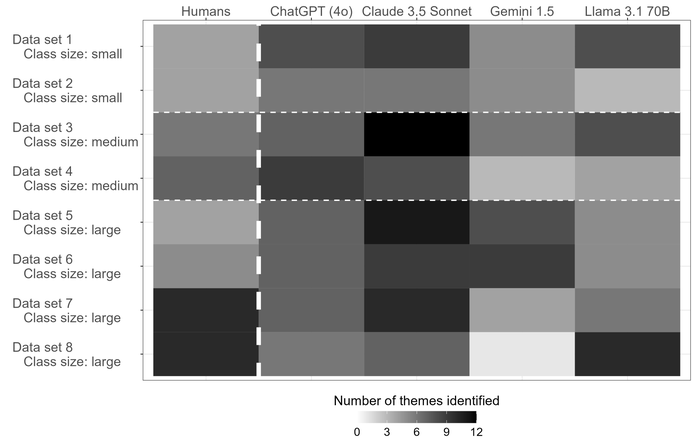
* Preliminary results reflect only half of the data and additional insights may emerge once remaining data is analyzed.

Georgia Physics and Astronomy Education Research
Franklin College of Arts and Sciences
UNIVERSITY OF GEORGIA

College of Veterinary Medicine
UNIVERSITY OF GEORGIA

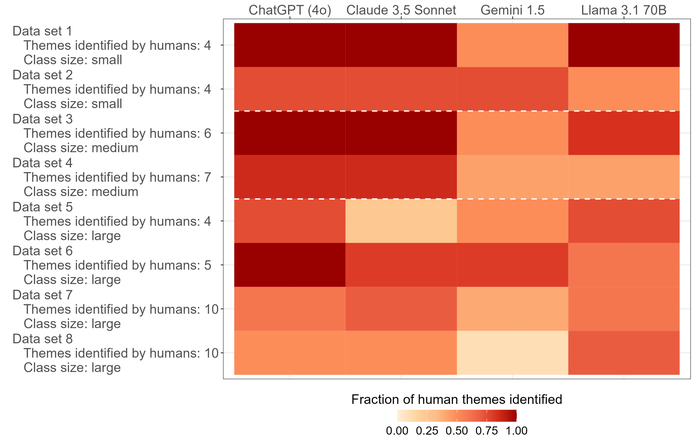
Preliminary result: Generative AI tools, such as ChatGPT and Claude, can extract themes from student feedback about as effectively as instructors.

How do models compare in their ability to find themes in the student feedback?



ChatGPT and Claude tended to find more themes in the course feedback than instructors did while Gemini and Llama tended to find fewer themes than instructors did.

Did the themes identified by AI tools align with what instructors identified?



For ChatGPT and Claude, yes. For Gemini and Llama, less so.

In many cases, ChatGPT and Claude identified all of the themes the instructors identified while this rarely happened for Gemini or Llama.

Of the themes identified by generative AI tools but not instructors, how many were hallucinations?

Analysis is ongoing but preliminary results suggest that many of the themes identified by generative AI tools but not instructors are not due to hallucinations but rather a result of only 1 or 2 instructors identifying the themes rather than 3 needed for a majority.

References

[1] Spooren, P., Brockx, B., & Mortelmans, D. (2013). *Review of Educational Research*, *83*(4), 598-642.
[2] Pu, X., Gao, M., & Wan, X. (2023). *arXiv preprint arXiv:2309.09558*.
[3] Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). *International journal of artificial intelligence in education*, 1-38.
[4] Based on https://ctl.uga.edu/teaching-resources/feedback-and-evaluation-of-teaching/interpreting-responding-to-student-evaluations-of-teaching/