Slides: https://bit.ly/

UNIVERSITY OF GEORGIA
Franklin College of Arts and Sciences

Large language models are effective for summarizing student feedback

Nicholas T. Young

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

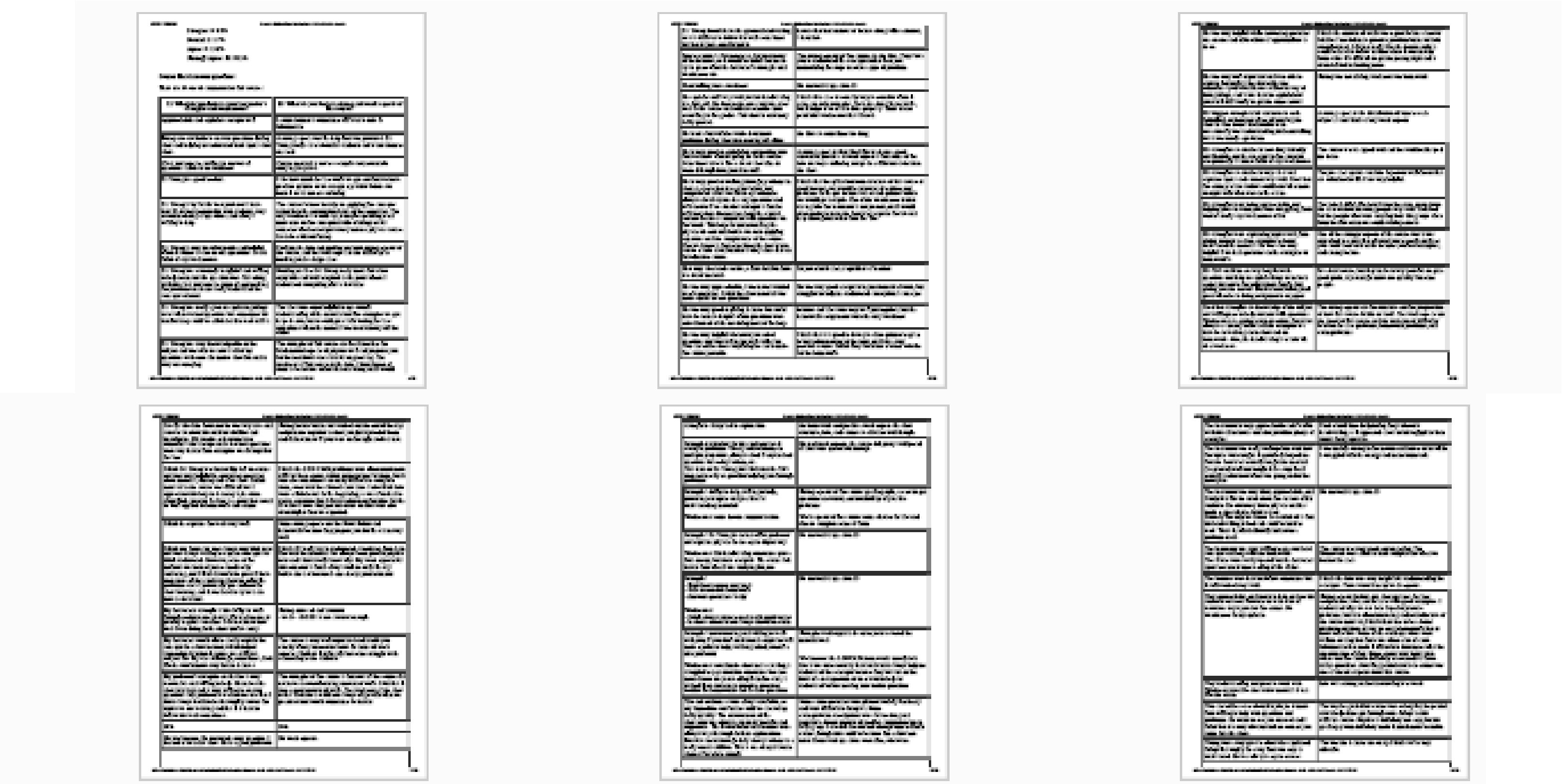
[1] Department of Physics and Astronomy, University of Georgia
[2] Department of Physiology and Pharmacology, College of Veterinary Medicine, University of Georgia
[3] Institute for Astronomy and Astrophysics, Eberhard Karls University of Tübingen

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

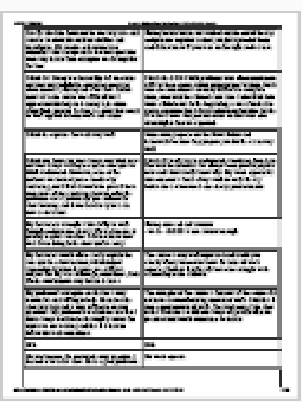
Slides

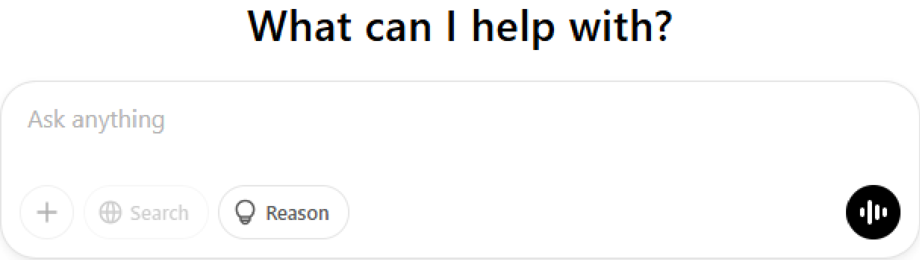
Student feedback is useful but can be time-consuming to understand

Student feedback is useful but can be time-consuming to understand



Chatbots can summarize text and could be a useful tool to quickly understanding student feedback





Pu et al 2023; Parker et al 2024

What’s the problem then?

What’s the problem then?

- Unstructured text

What’s the problem then?

- Unstructured text
- Sample size

What’s the problem then?

- Unstructured text
- Sample size
- Hallucinations

Xu et al 2025

Determine how well LLMs can extract key insights from student feedback

Data

- Student responses to two questions on end-of-course feedback survey
  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?*
- 9 courses taught by 3 unique instructors

Methodology

- 5 instructors read each set of anonymized student feedback and create a summary
- Same feedback files shared with 4 AI tools
  - LLMs have inherent randomness so did 5 trials with each model

Methodology

- Prompt: “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”

https://ctl.uga.edu/teaching-resources/feedback-and-evaluation-of-teaching/interpreting-responding-to-student-evaluations-of-teaching/

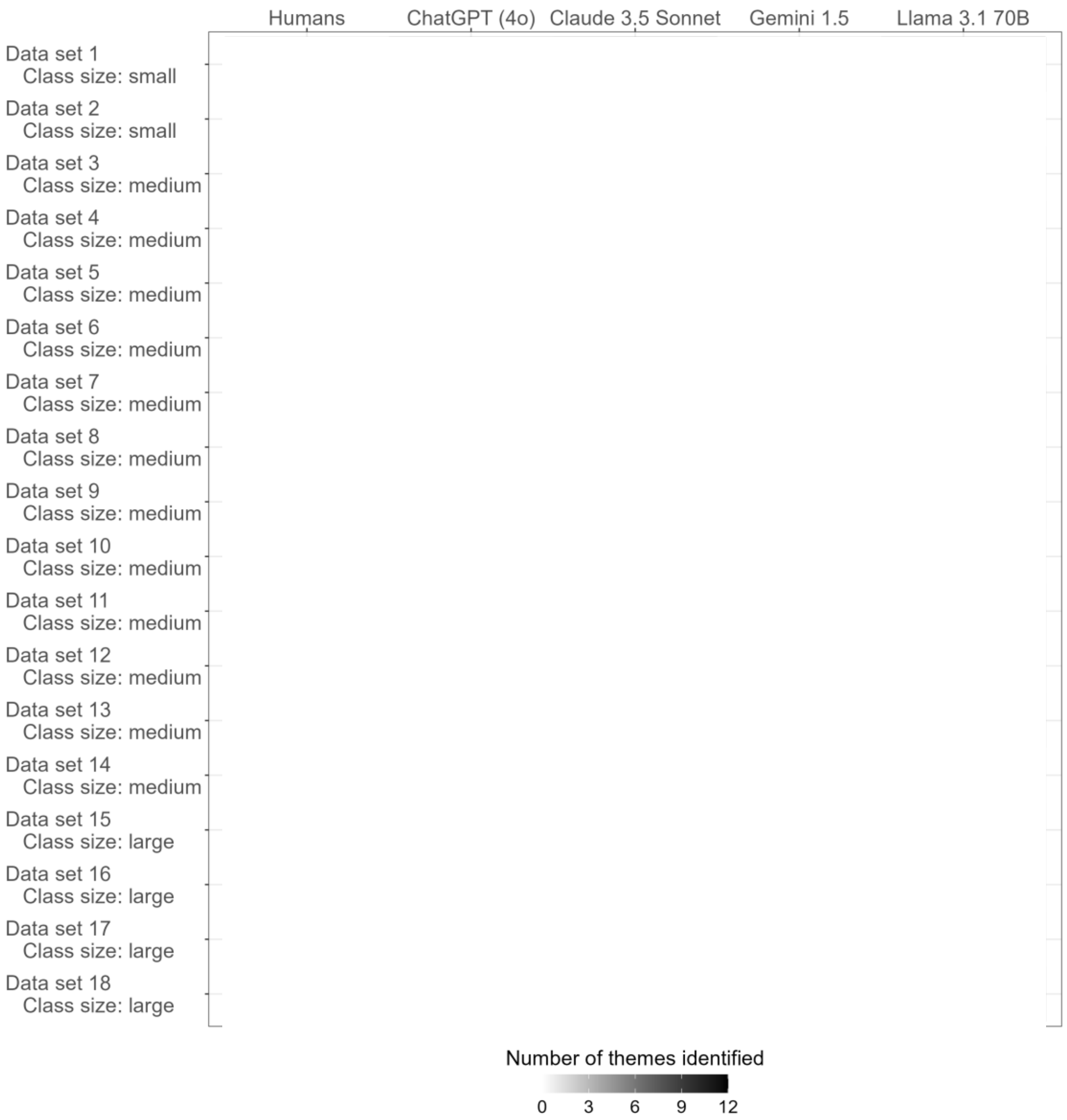
Analysis

- Majority voting (3 of 5) to be included in final summary for each tool and 3 of 5 instructors

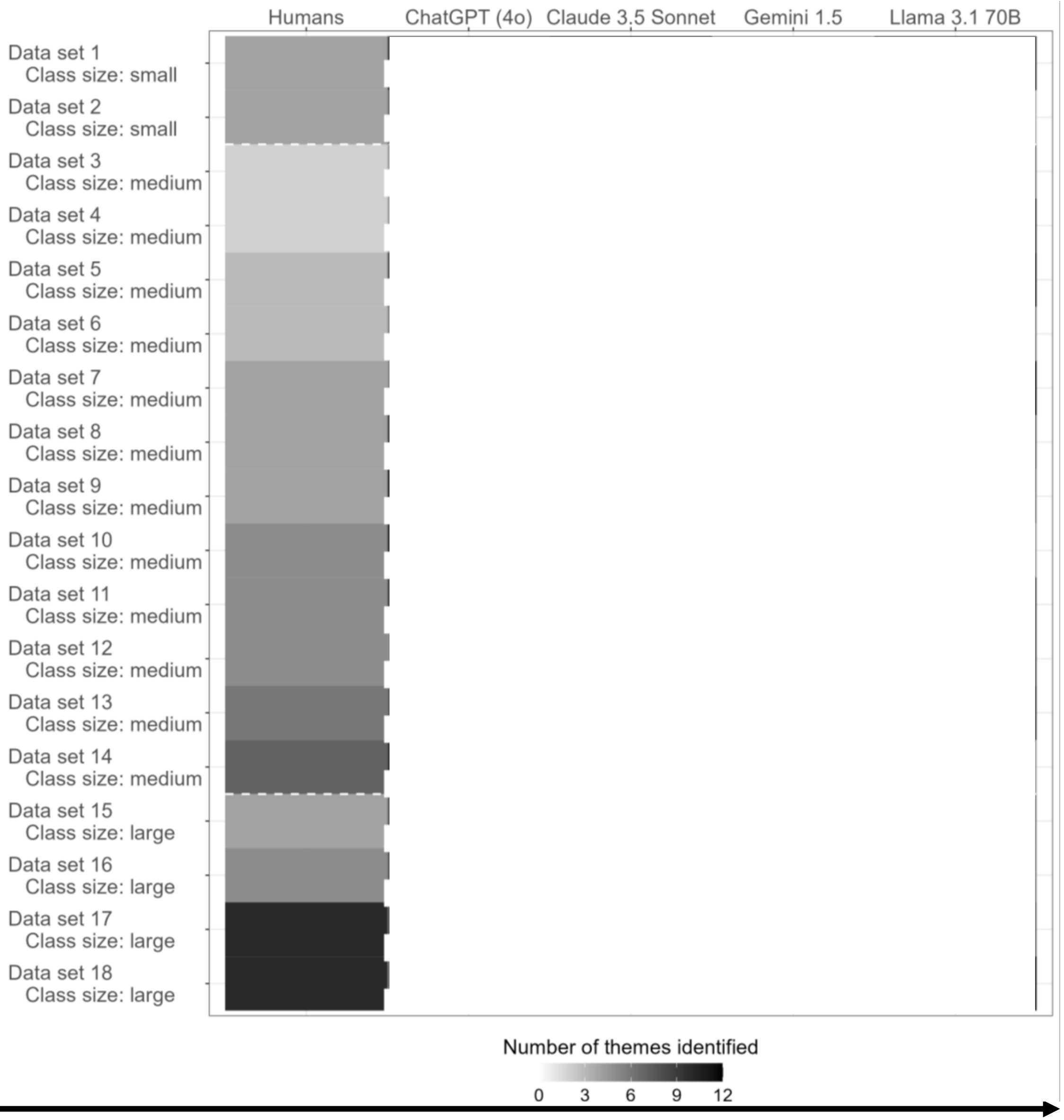
Analysis

- Majority voting (3 of 5) to be included in final summary for each tool and 3 of 5 instructors
- Take instructor summary as “true” summary and compare LLM summaries with human summaries

Results: themes identified by group

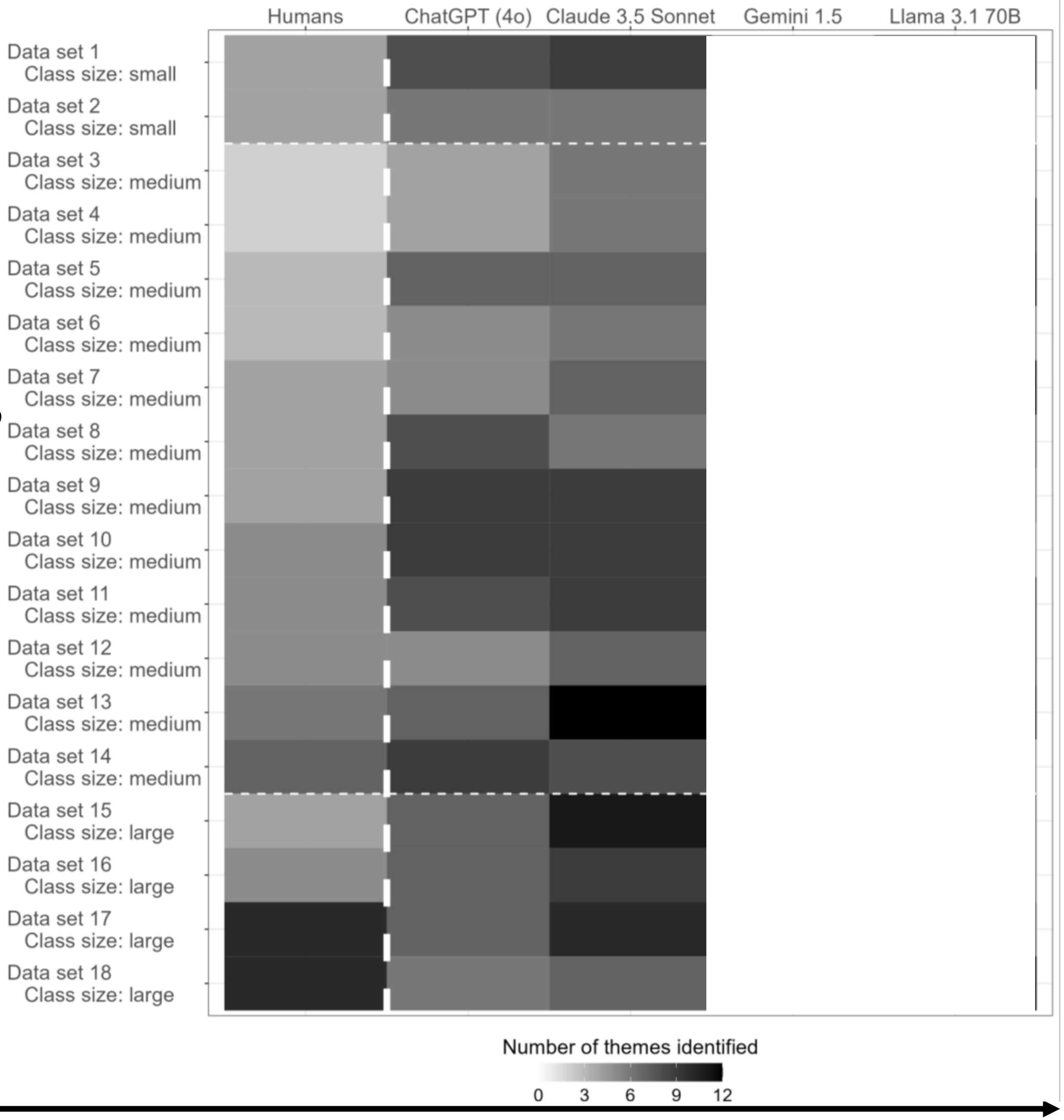


Results: Themes identified by model

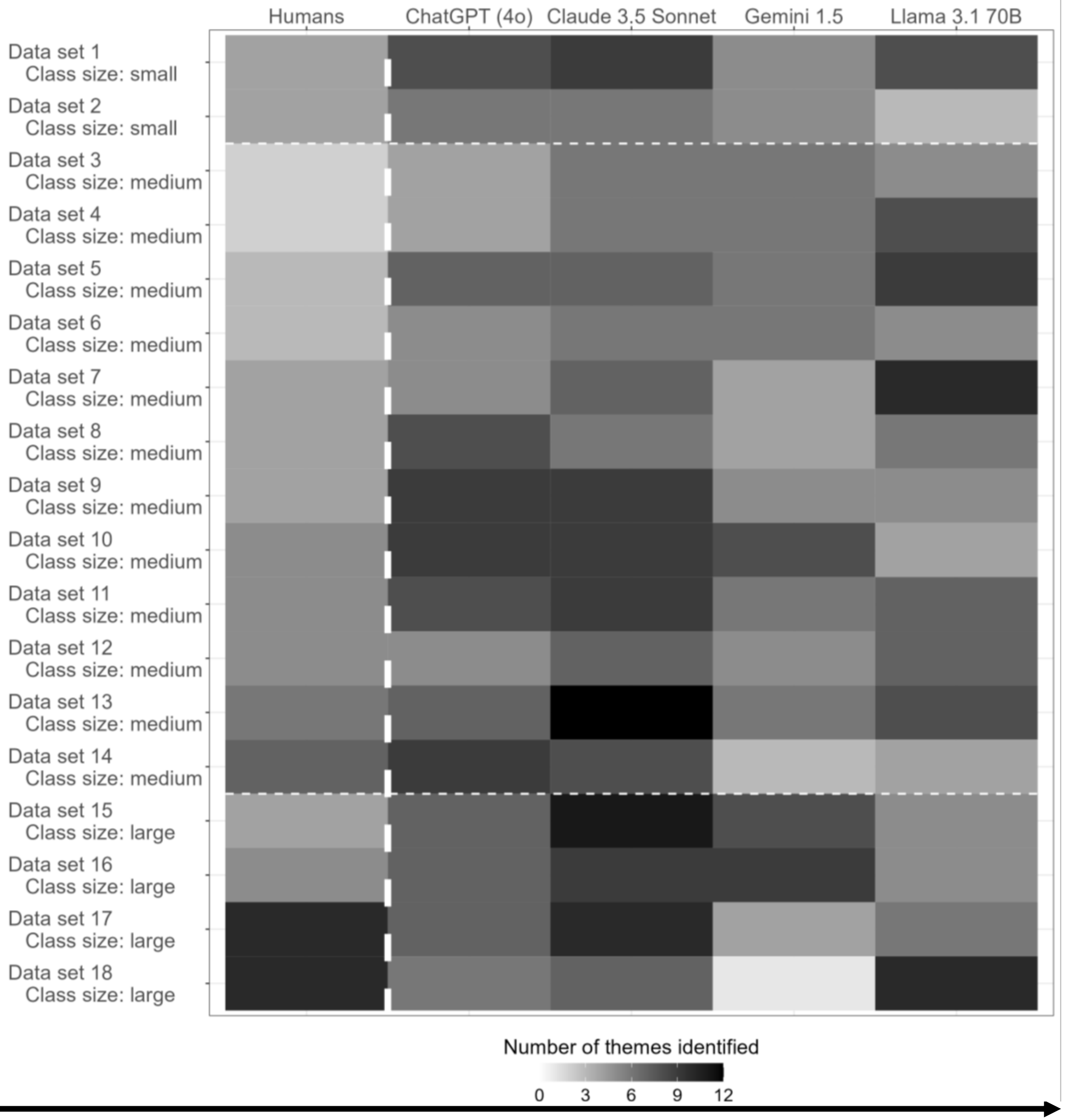


Nicholas Young

Results: ChatGPT and Claude identify more themes than instructors do

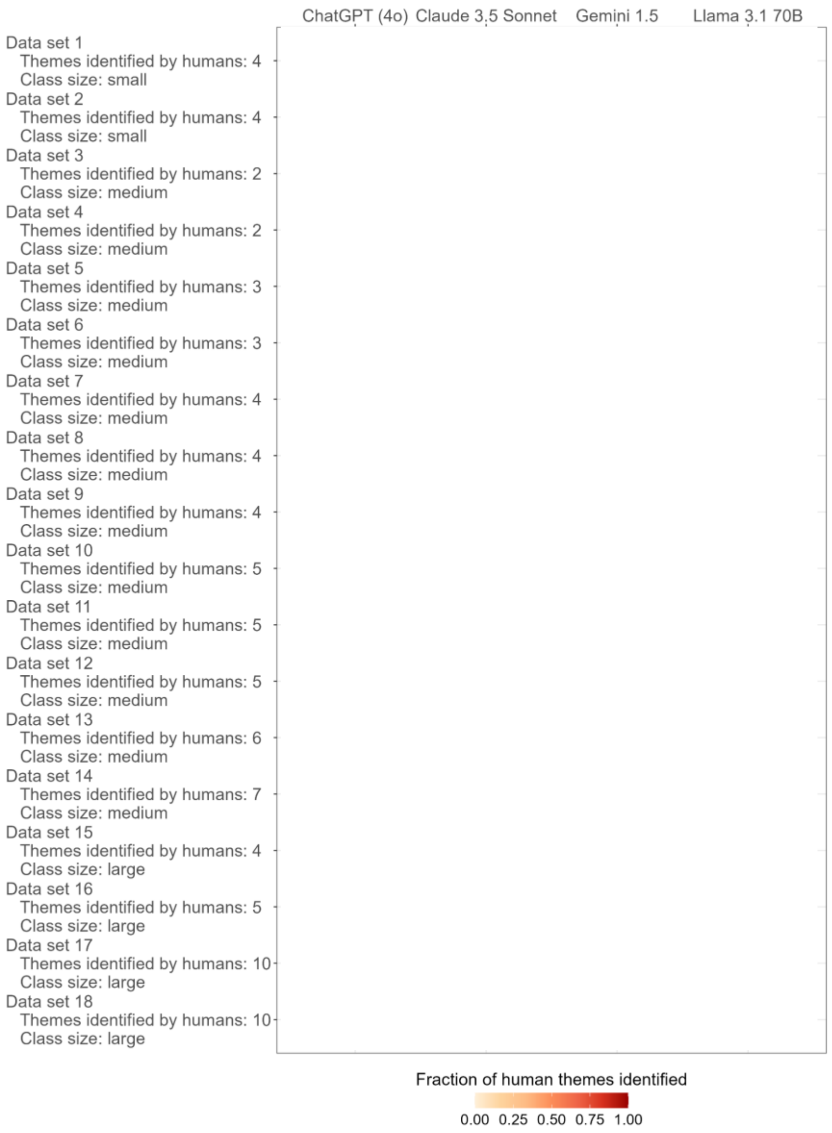


Results: Gemini and Llama have mixed results

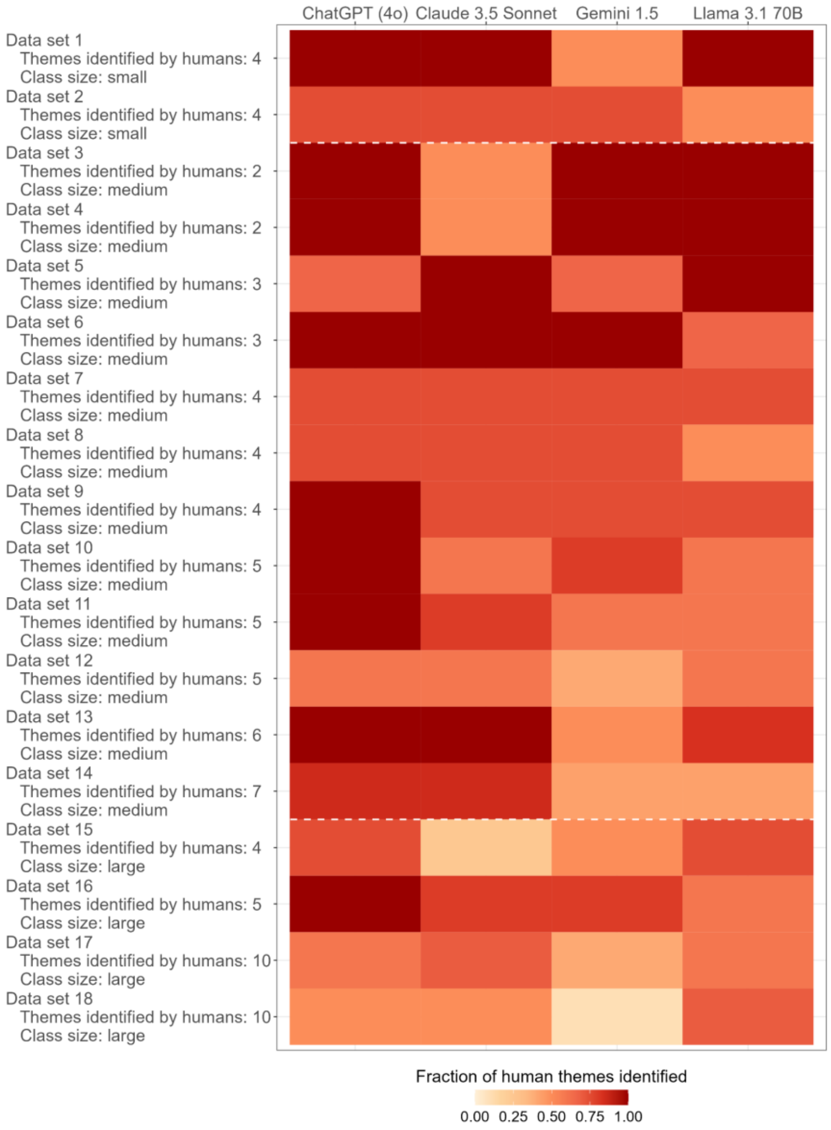


How do the LLMs do in comparison to the instructors?

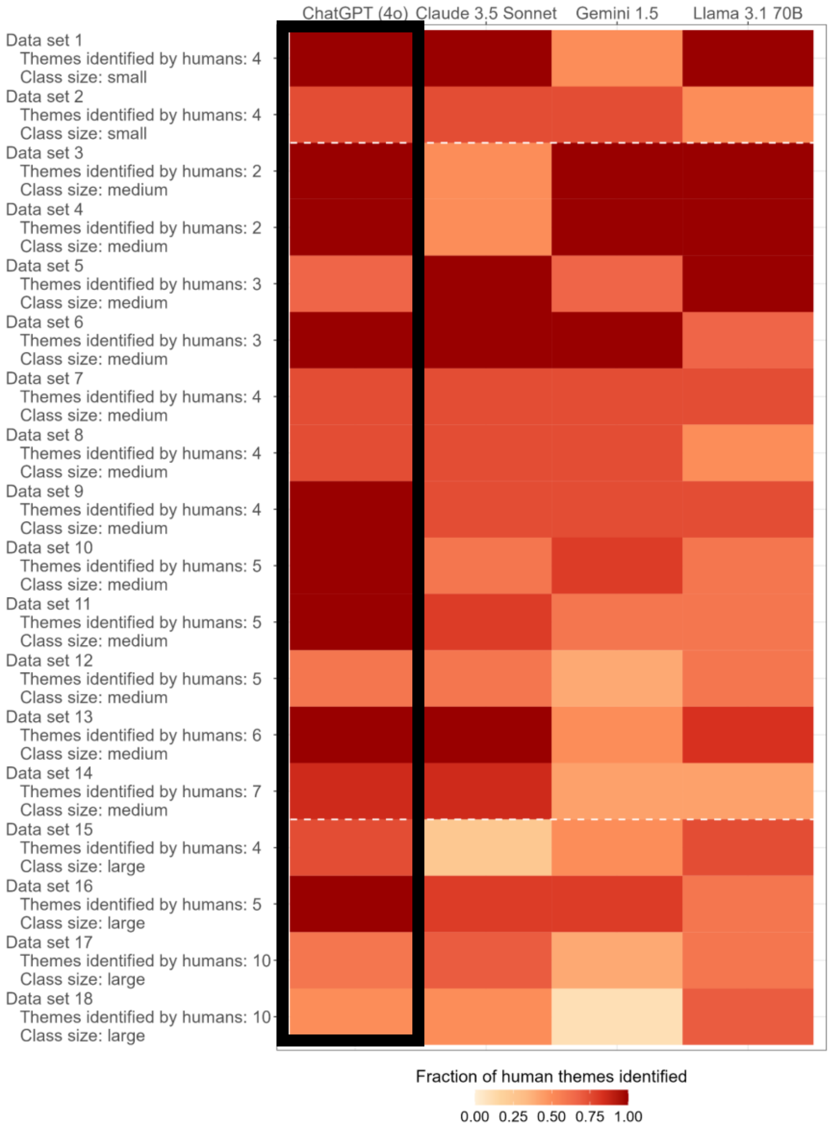
Results:


ChatGPT (4o)
Claude 3.5 Sonnet
Gemini 1.5
Llama 3.1 70B
Data set 1
Themes identified by humans: 4
Class size: small
Data set 2
Themes identified by humans: 4
Class size: small
Data set 3
Themes identified by humans: 2
Class size: medium
Data set 4
Themes identified by humans: 2
Class size: medium
Data set 5
Themes identified by humans: 3
Class size: medium
Data set 6
Themes identified by humans: 3
Class size: medium
Data set 7
Themes identified by humans: 4
Class size: medium
Data set 8
Themes identified by humans: 4
Class size: medium
Data set 9
Themes identified by humans: 4
Class size: medium
Data set 10
Themes identified by humans: 5
Class size: medium
Data set 11
Themes identified by humans: 5
Class size: medium
Data set 12
Themes identified by humans: 5
Class size: medium
Data set 13
Themes identified by humans: 6
Class size: medium
Data set 14
Themes identified by humans: 7
Class size: medium
Data set 15
Themes identified by humans: 4
Class size: large
Data set 16
Themes identified by humans: 5
Class size: large
Data set 17
Themes identified by humans: 10
Class size: large
Data set 18
Themes identified by humans: 10
Class size: large
Fraction of human themes identified
0.00
0.25
0.50
0.75
1.00

Results: All models are generally better at finding the same themes as humans when humans find fewer themes

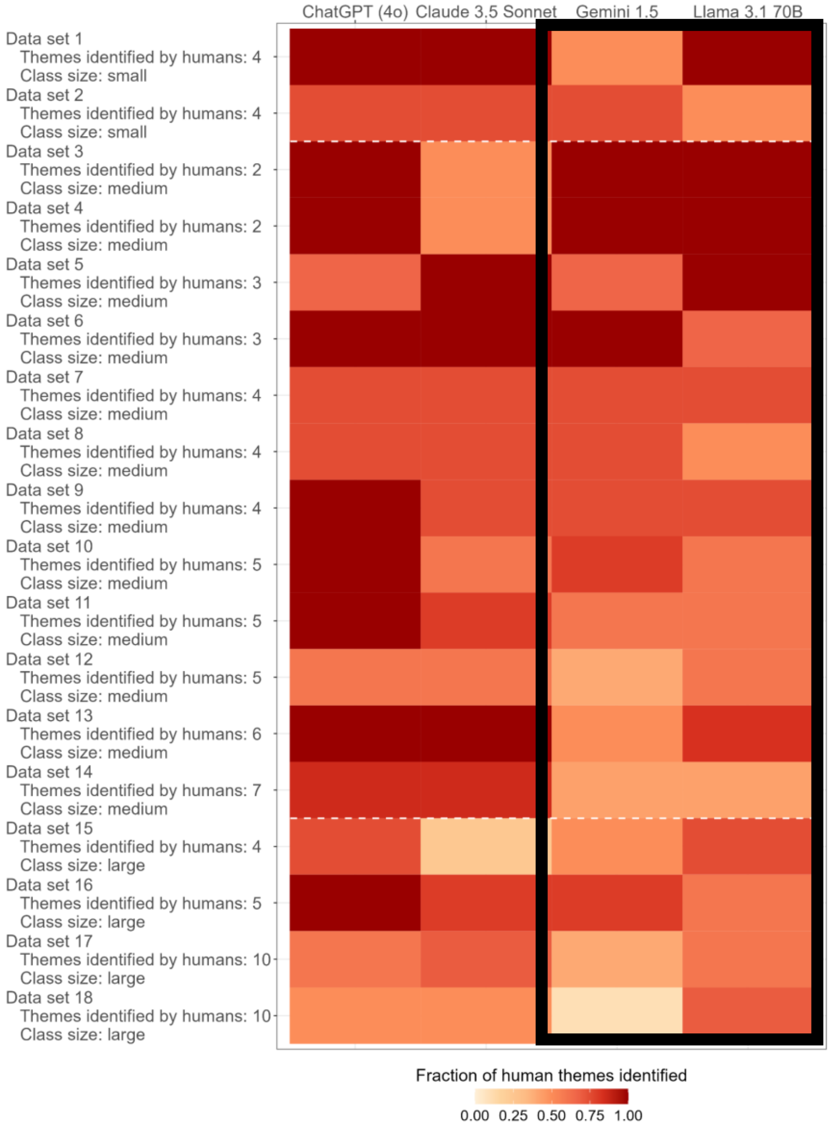


Results: ChatGPT is generally best at finding the same themes as humans

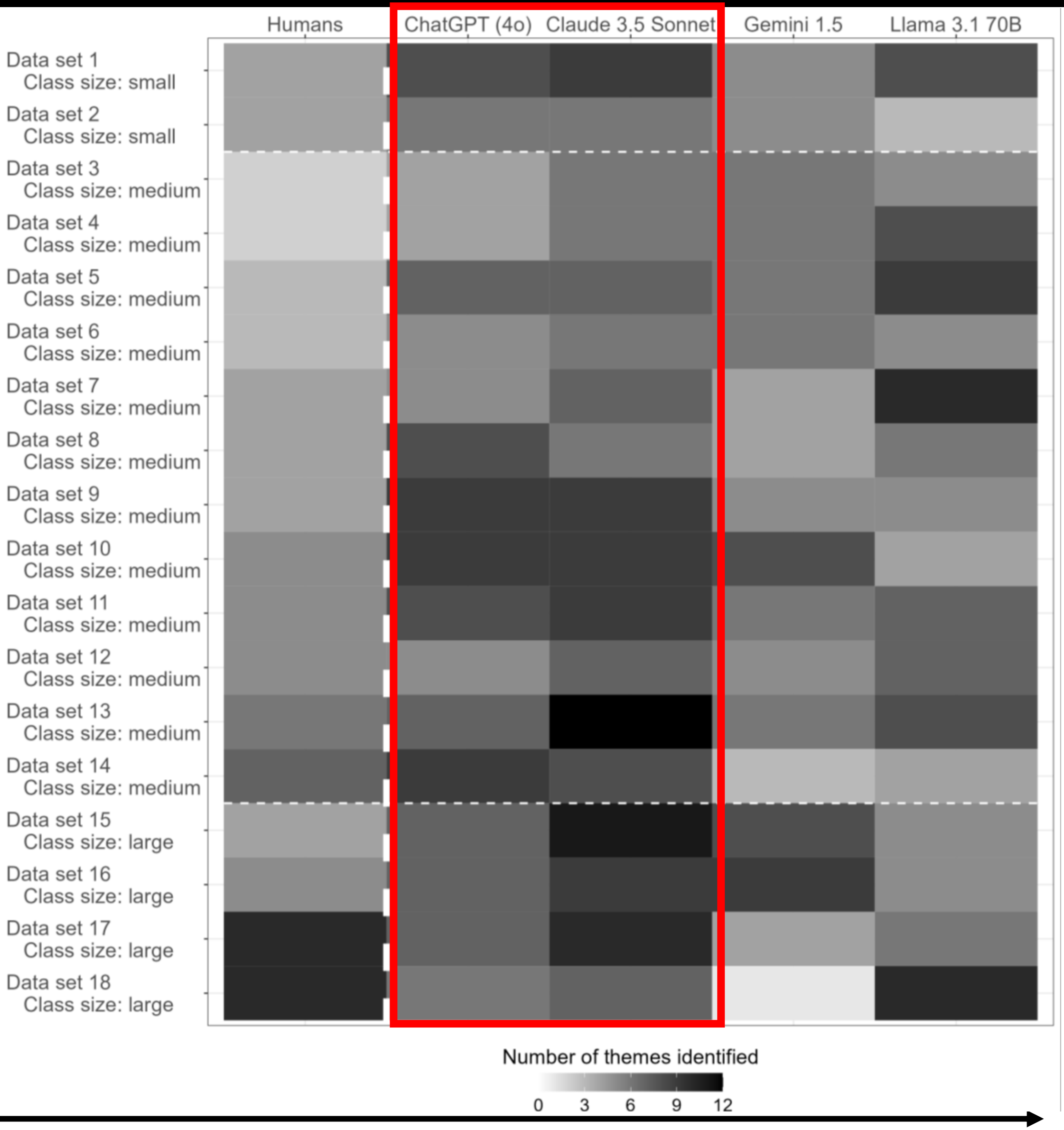


Nicholas Young

Results: Gemini and Llama were generally the worst at finding the human-identified themes



What about the themes not identified by humans?



Nicholas Young

Are the extra themes hallucinations?

- Analysis still in progress

Are the extra themes hallucinations?

- Analysis still in progress
- Probably not
    - Likely result of only 1 or 2 instructors identifying theme

Takeaways

- LLMs can be useful for summarizing student feedback

Takeaways

- LLMs can be useful for summarizing student feedback
- ChatGPT showed the best performance followed by Claude, then Llama and Gemini

Takeaways

- LLMs can be useful for summarizing student feedback
- ChatGPT showed the best performance followed by Claude, then Llama and Gemini
- Instructors likely want to ask for a summary multiple times or spot check results

Takeaways

- LLMs can be useful for summarizing student feedback
- ChatGPT showed the best performance followed by Claude, then Llama and Gemini
- Instructors likely want to ask for a summary multiple times or spot check results

Questions?

nicholas.young@uga.edu