Applying Natural Language Processing with Fine-Tuned Large Language Models to Streamline Student Survey Data in Large Enrollment Courses

Georgia Tech College of Sciences
School of Mathematics

Adi Krish (adikrish@gatech.edu), Greg Mayer (greg.mayer@gatech.edu)
Georgia Institute of Technology

GitHub

Project Website

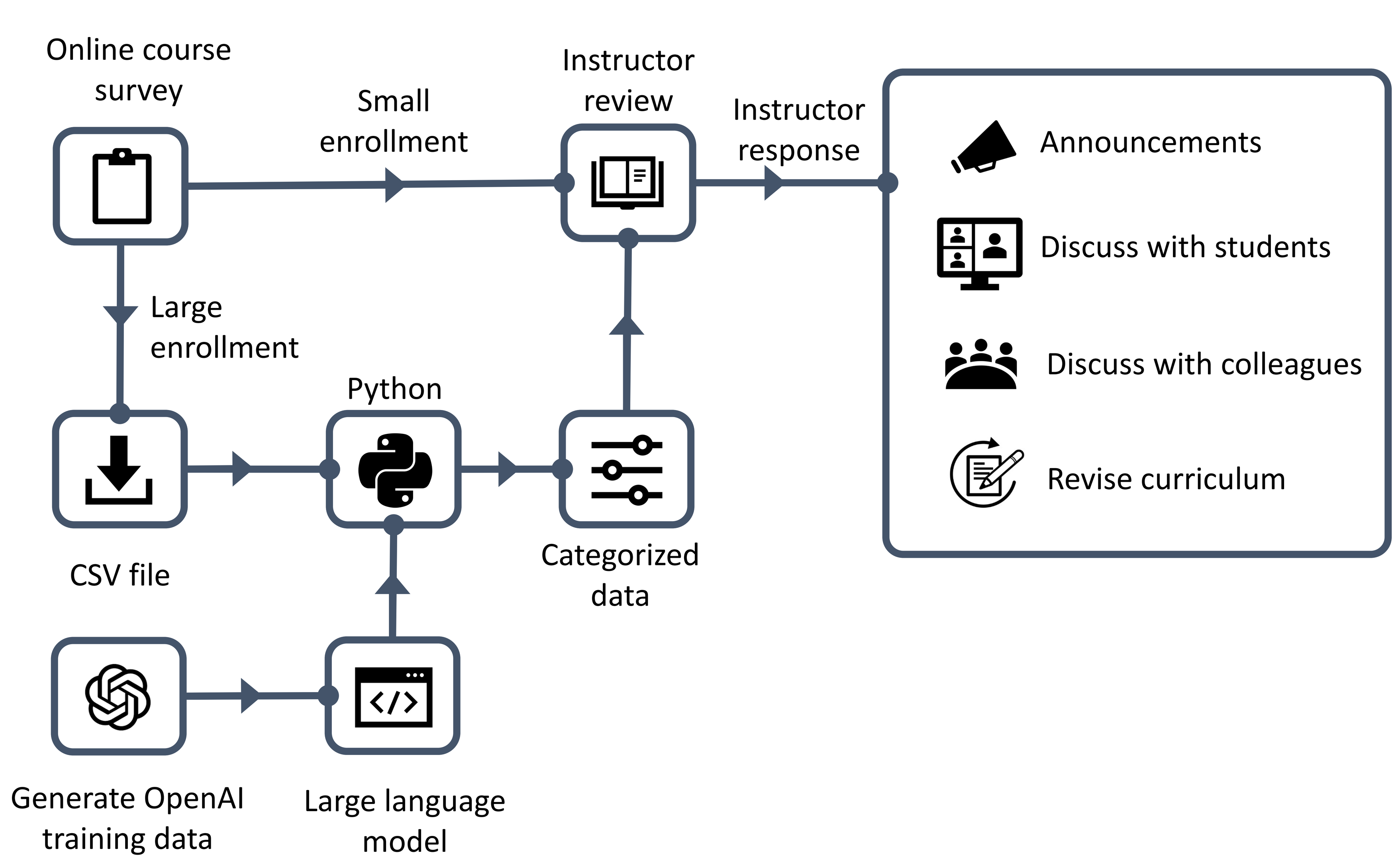
Motivation

- Start-of-semester and mid-semester surveys help instructors become more aware of student needs[1,2].
- We have found that interpreting survey data can be a challenge when:
  - Surveys include open-response questions.
  - Enrollment of the course is large (> 500 students).
- Open-response data can have items that the instructor may need to identify but can be missed as the size of the data increases. Specific examples:
  - Requests for support.
  - Questions for the instructor.
  - Concerns and/or needs that the instructor may be able to address.

Methods

- Machine learning algorithms offer many methods to assist with text classification[3].
- Our approach augments the approach that an instructor might use without AI.
- OpenAI reduces need for human annotation with a large language model (LLM)[4].
- **Requirements**
  - FERPA and institute privacy policy compliance.
  - Reduce time instructor uses to process survey data.
  - Instructor must read all student responses.
- **Hypothesis**
  - Categorizing survey responses reduces time needed to process data.
- **Assumptions**
  - Instructor will review all responses after categorization for accuracy.
  - Survey data can be grouped into categories useful to instructor.

Instructor Workflow

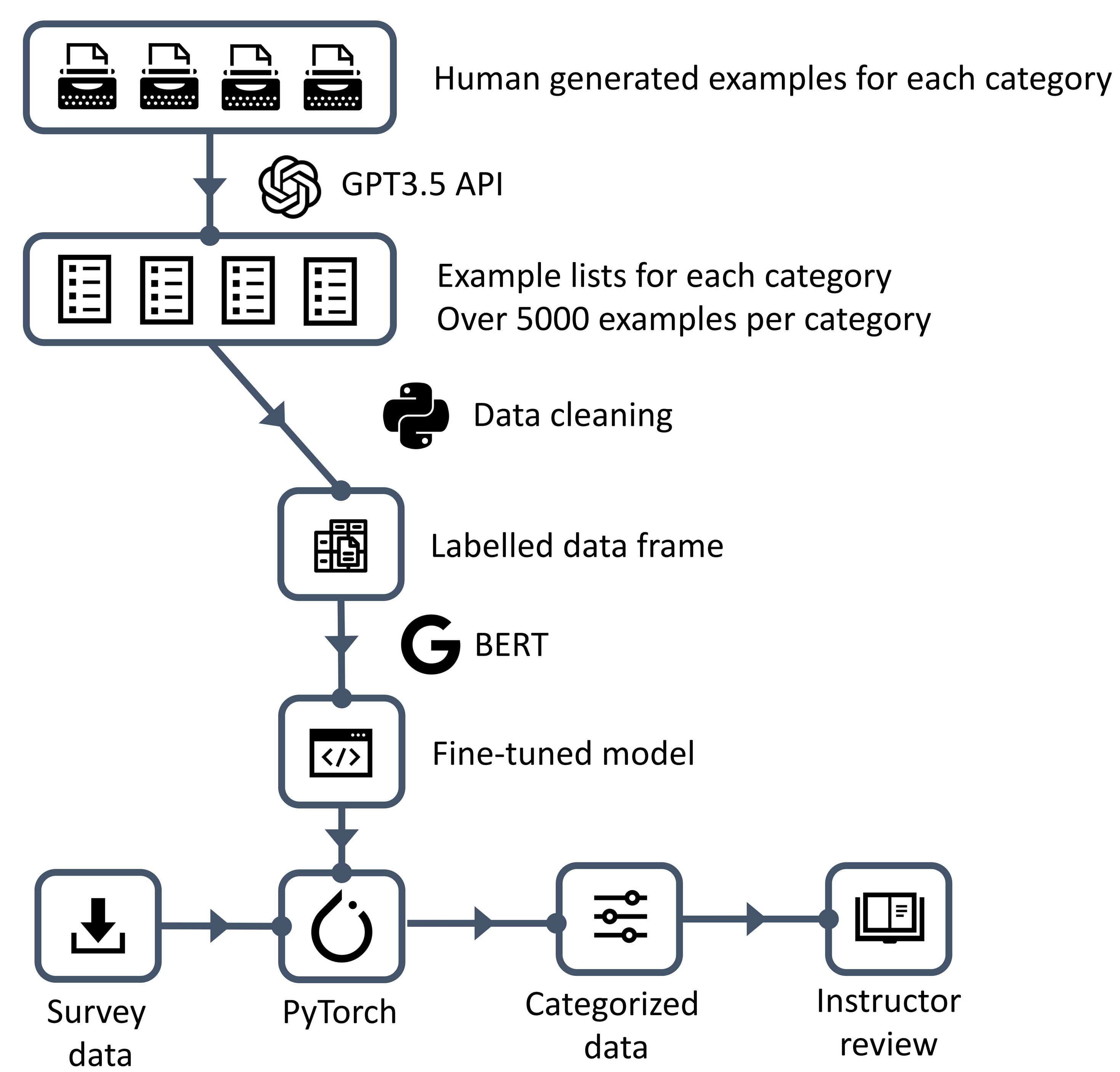


Data Categories

- Start-of-semester survey data hypothesized to have four useful categories for identification, defined below.
- Categories defined based on how instructor might respond to student.

Code	Definition	Action
NC	No concerns	Thank student for completing survey
LM	Learning management concern (e.g. exam stress, reviewing pre-req)	Direct student to recommend resources
TM	Time management concern	Direct student to recommend resources
OT	Other (i.e. – all other comments)	Case-by-case

Data Classification Process



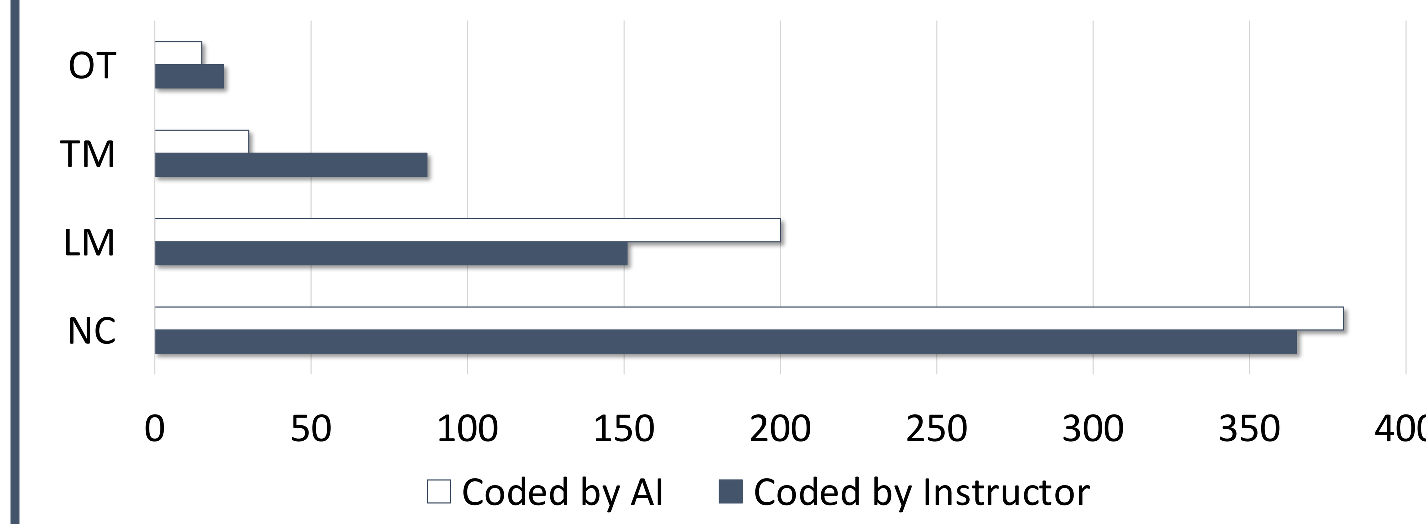
Methods

- Applied our classification process to a start-of-semester survey data set.
- Data had roughly 600 responses from a large course taught at GT.
- Analyzed data from an open response question: *At this point in the semester, what concerns you the most about taking this course, if any? If you do not have any, you can write "none" or "NA".*
- Open response survey data was:
  1. Human coded by instructor using four categories (NC, LM, TM, OT).
  2. Machine coded by our Augmented Data Fine-Tuned BERT.

Preliminary Results

- Overall percent agreement across all categories was roughly 83%.
- Concerns (OT, TM, and LM):
  - Instructor coded 260 comments as concerns.
  - Model coded 240 comments as concerns (92.3% match).
- No concern (NC):
  - Instructor coded 365 comments as no concern.
  - Model coded 360 comments as no concern (98.63% match).

Start-of-Semester Survey Code Distribution



Analysis and Conclusions

- Model struggled with complex statements. For example:
  - Statements that included multiple concerns.
  - Statements that expressed a concern and a no-concern sentiment (eg – I had a concern, but I am ok now).
- Instructor would still need to review all comments so that no concerns would be missed.

Future Work

- Apply methods to mid-semester survey data and other courses.
- Reduce number of categories to decrease categorization errors.
- Detach from GPT API to make model free (i.e. – no cost).
- Develop methods to fine tune with real/complicated data on top of existing fine tuning.
- Experiment with different categories, for example:
  - Algorithms that generate categories rather than relying on predefined categorizes.
  - Only two categories: no concerns vs everything else.

References

1. Richardson, J. T. (2005). Instruments for obtaining student feedback: A review of the literature. Assessment & evaluation in higher education, 30(4), 387-415.
2. Angelo, T. A., & Cross, K. P. (2012). Classroom assessment techniques. Jossey Bass Wiley.
3. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.
4. Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., ... & Li, X. (2023). Auggpt: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007.