How well does ChatGPT perform on course assignments from the TU Delft Computer science and engineering Bachelor?

Mike Segers M.Segers-1@student.tudelft.nl

Supervisor: Xiaoling Zhang, Fenia Aivaloglou

Introduction

- Research capabilities of AI
- Creating academic enviroment for collaborating with AI

Methods

- **Gathering data:** Data collected through a combination of methods (material contribution and Brightspace page scraping) from various bachelor CSE courses
- **Large language model:** ChatGPT based on GPT version 3.5
- **Prompt engineering:** Original questions, with minor edits for copy errors. One retry to correct wrong answer, with no additional information.
- **Evaluation of results:** MCQs were objectively categorized, while open-ended questions were manually checked

Dataset

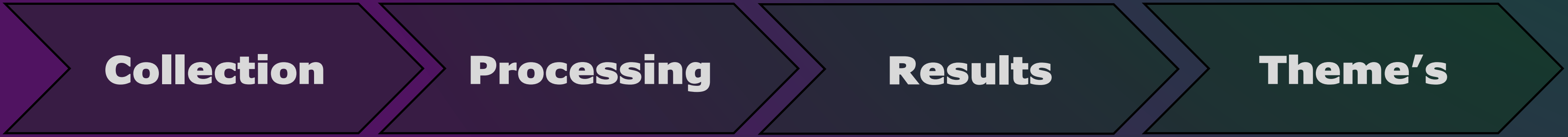
- 6 courses
- 349 MCQs
- 215 open questions

Results

- 64% average MCQ score
- 41% average open question score
- 52% MCQ improvement after retry
- 12% open question improvement after retry

Limitations

- Courses who approved
- single study and university
- Older version of GPT
- No prompting technique
- Model limitations:
  ▣ Exercises with images
  ▣ MCQ: Often right reasoning but wrong choice



Collection

Processing

Results

Theme’s

Discussion

- High performance
  ▣ Definitions & concepts
  ▣ Connections between topics
- Low performance
  ▣ Large code understanding
  ▣ Complex mathematical reasoning

Most relevant work

- D. Nunes, R. Primi, R. Pires, R. Lotufo, and R. Nogueira. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams, 2023.
- J. Savelka, A. Agarwal, C. Bogart, Y. Song, and M. Sakr. Can generative pre-trained transformers (gpt)pass assessments in higher education programming courses?, 2023.