

A Custom Generative AI Chatbot as a Course Resource

Yutong Ai, Maya Baveja, Akanksha Girdhar, Melina O'Dell,
Andrew DeOrio

University of Michigan

Problem

- The use of Generative AI has increased quickly
 - Examples: ChatGPT, Khanmigo
- Can Generative AI be helpful as a course resource?
 - In addition to office hours and course forums

Research questions

RQ1: Can a customized GenAI chatbot be helpful to students as a course resource?

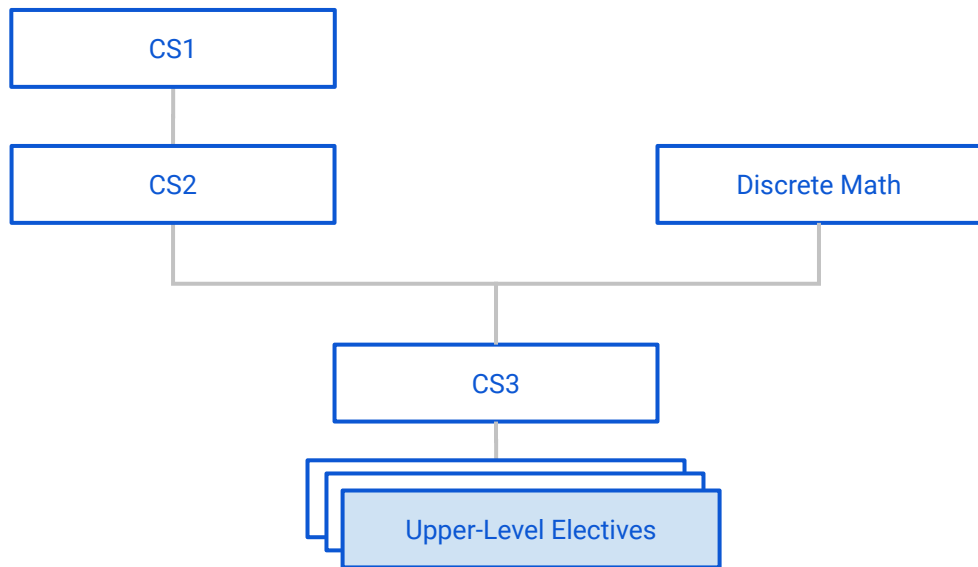
RQ2: Is hallucination (false or misleading responses) a barrier to student learning?

RQ3: How do customized chatbot responses compare to a general purpose chatbot?

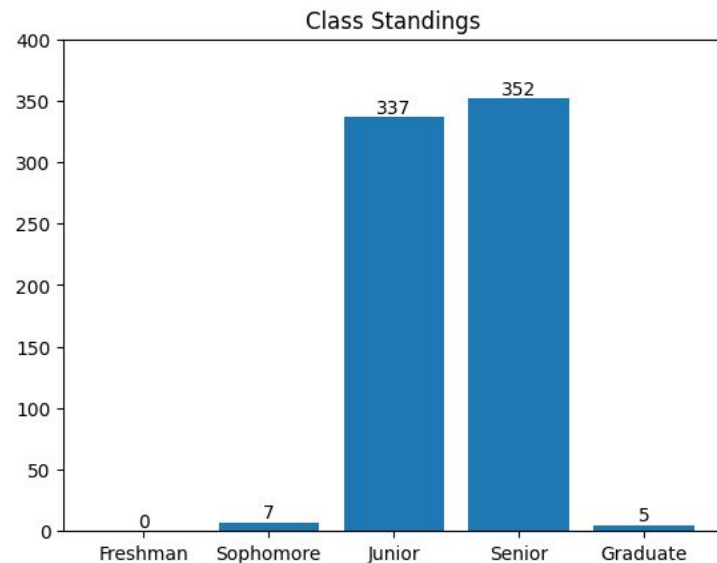
Methods

- Released a custom chatbot to students, trained on assignment materials
- Surveyed students to collect general feedback about the Bot and a sample of prompt/response interactions
- Analyzed hallucinations, helpfulness, and customization metrics from survey data using an evaluation by team of experts and comparison with a general purpose chatbot

Course and project



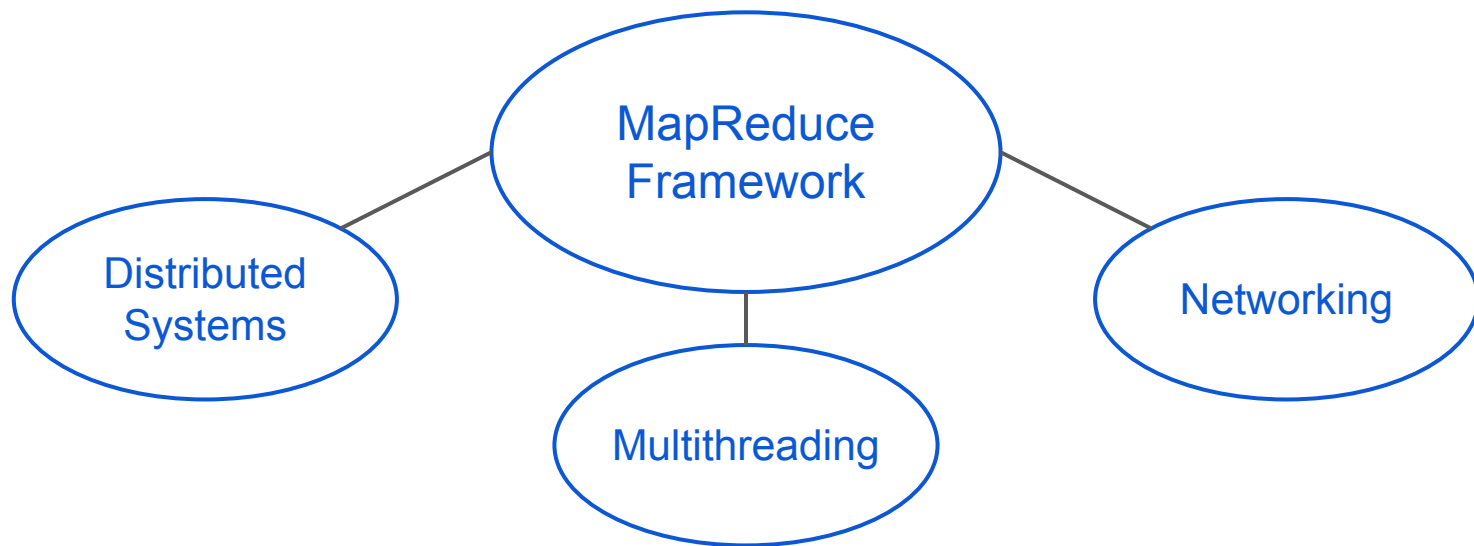
Course: Upper-level CS course focusing on web systems, with over 700 students



Most students are juniors and seniors

Course and project

Most students worked in groups of 2-3 and had no prior experience with either networking or parallel programming



The custom Bot

- ChatGPT-like interface for submitting prompts



Can you explain how the Manager handles fault tolerance?



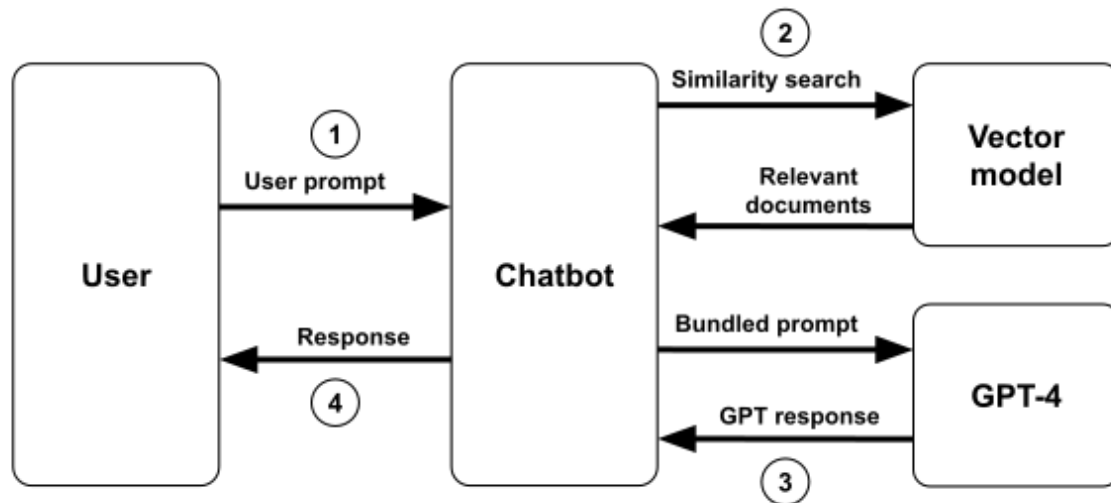
Indeed, in the web development scenario outlined, the Manager is responsible for handling tasks in the presence of failure by various means of fault tolerance procedures. From what has been provided:



If a worker (W_0) fails during a task assignment, the Manager requeues the task (T_0) that was supposed to be executed by the now defunct worker. In other words, from the Manager's perspective, any task that was in the process of assignment when a failure is detected simply goes back into a queue for assignment to another worker.

Bot model

- Based on OpenAI's GPT-4
- Retrieval Augmented Generation (RAG) system:
 - Searches for 4 most similar documents
 - Bundles with prompt
 - Prompts LLM



Training data

The Bot was trained on:

- Project specification (HTML)
- 5 Instructor-written tutorials (HTML)
- Slides from 2 relevant labs (PDF)
- Slides and transcripts from 4 relevant lectures (PDF, TXT)
- Question-and-answer threads from relevant course forum posts over 3 semesters (TXT)

Configuration

- Created a teaching assistant "persona" with system prompt

“Imagine you are a helpful teaching assistant for a web development course. Use the following pieces of context to answer the question at the end. If you don’t know the answer, just say that you don’t know, don’t try to make up an answer. {context} Question: {question} Helpful Answer:”

- Used temperature 1.2 (out of 2) to maximize quantity of helpful responses

Survey

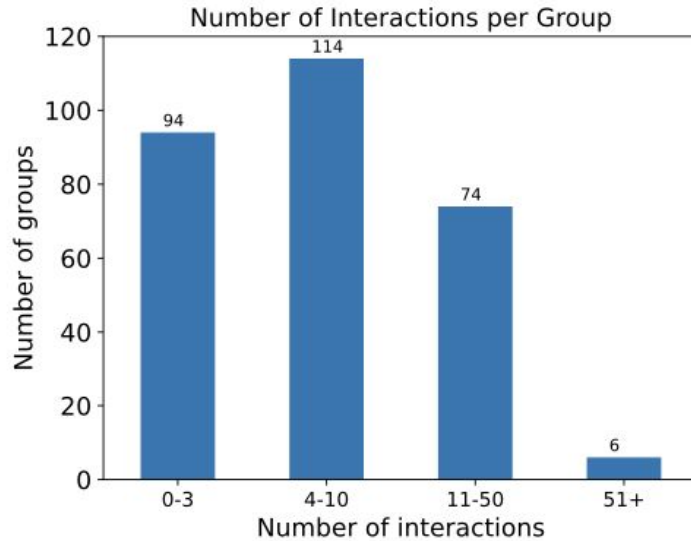
- **Background:** Student class standing and prior experience level with project topics
- **General:** Overall helpfulness of the Bot, number of interactions, its use relative to traditional course resources, and whether the Bot improved coding performance and/or saved time
- **Sample interactions:** 3 example student-Bot interactions, consisting of student prompt, Bot response, and whether they thought the response was correct and/or helpful

Results – Bot helpfulness

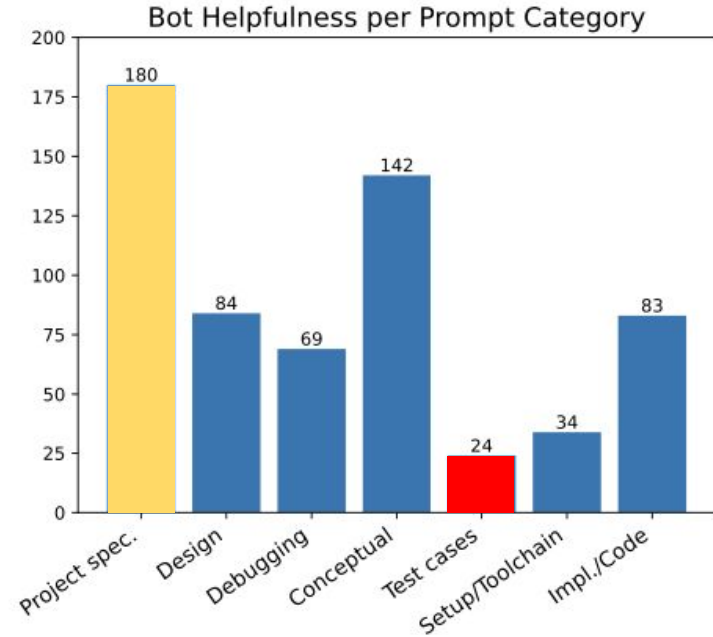
- 289 valid survey responses containing 834 valid student-Bot interactions
- 77.0% of interactions were helpful, 60.9% of survey responses said that a similar bot would be helpful for other projects
- Upon expert instructor review the majority of student-reported "unhelpful" responses were actually helpful
 - Responses from requests for solution code
 - Responses from low-quality prompts (i.e. vague or obscure questions)

A custom chatbot is a helpful resource for a specific course project
But student evaluations of helpfulness may reflect expectations of the Bot

Results – Bot helpfulness



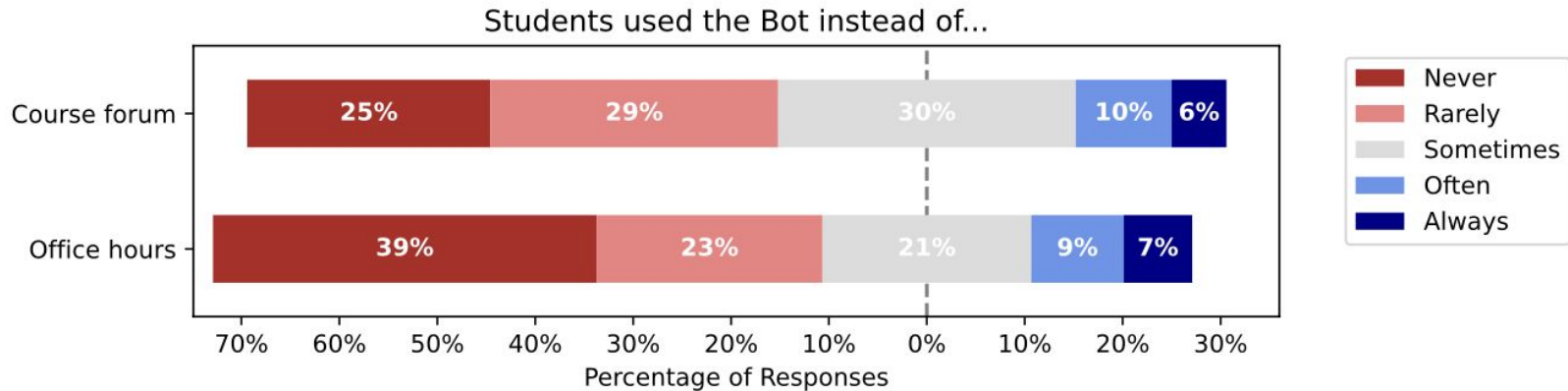
Number of interactions with the Bot for each student group. Most student groups used the Bot at least 4-10 times.



Student-reported Bot helpfulness per prompt category. Students found the Bot to be most helpful for project specification questions.

Results – Comparison with other course resources

- Most students did not use the Bot instead of office hours or the course forum



The Bot may augment (but not replace) traditional course resources

Results – Student prior experience

- Did not observe any statistically significant associations between:
 - Class standing or prior parallel/networking programming experience
 - Number of interactions with the Bot and whether the Bot saved time

Results – Hallucination

- Team of expert instructors evaluated every Bot response for correctness and helpfulness using the student prompt
- **92% of Bot responses were correct, 85% were helpful for completing the project**
 - Students correctly identified 96% of correct responses as correct
 - Students correctly identified 94% of helpful responses as helpful
- **96.3% of the incorrect responses were correctly identified by students**

Hallucination was not a barrier to student learning

Results – Customized vs. general purpose chatbot

- Prompted our Bot and a generic state-of-the-art chatbot (ChatGPT) with sample student prompts
- Bot responses were highly **detailed and specific**
- ChatGPT responses were vague or did not align with project requirements

Prompt Category	Our Bot	ChatGPT
Project specification	✓	×
Design	✓	×
Debugging	✓	✓
Conceptual	✓	✓
Explaining test cases	✓	×
Setup/Toolchain	✓	✓
Implementation/Code suggestions	✓	*

✓	Helpful responses
×	Unhelpful responses
*	Helpful for general prompts, unhelpful for project-specific prompts

The custom concentrated scope Bot was more helpful than a general purpose chatbot in the context of a specific course project

Limitations

Student Feedback

- Some students did not fill it out conscientiously
- The influence of completion credit may affect sincerity of their responses
- Student interactions may not be representative (only reported 3 each)
- No baseline for answering comparative survey questions
 - Ex. “Did responses help you code faster?”

Bot Constraints

- Lack of conversational responses and code files as training data due to software constraints
- Scope of the study is confined to one project within one upper-level course

Conclusions

We deployed a custom GenAI chatbot to assist with a distributed systems project

We found:

- The customized Bot provided overall helpful and correct responses and was helpful in assisting students with the project (RQ1).
- Hallucination was not a barrier since it was rare and easily identifiable (RQ2).
- Our Bot had advantages due to its concentrated scope (RQ3).

Our results can provide insights for faculty considering a customized chatbot as a course resource