Specialist vs. Generalist Generative AI for Student Learning

Melina O’Dell and Andrew DeOrio
University of Michigan

Context

- Generative AI tools are popular among students and educators
- **Goal:** Make a custom chatbot that uses course resources to help students on assignments
- How? **Retrieval-augmented generative AI (RAG):**
    - Pulls from external knowledge base to provide additional context for user prompt to LLM
    - Known to perform better than general-purpose chatbots (ex. ChatGPT) on domain-specific tasks by focusing LLM output

Problem

- Do RAG chatbots help students on programming assignments more or less than a general-purpose chatbot?

- What scope of external data is best for a course help RAG chatbot?
  - Course-wide → Less focused, worse performance?
  - Assignment-specific → More specific, too much effort for courses with many assignments?

Research Questions

RQ1: Do hallucination rates differ between *specialist* and *generalist* chatbots?

RQ2: Do helpfulness rates differ between *specialist* and *generalist* chatbots?

RQ3: Does the scope of a *specialist* chatbot affect hallucination or helpfulness?

Specialist: Custom RAG chatbots supplied with course materials

Generalist: General-purpose ChatGPT trained on ≈ entire internet

Methods

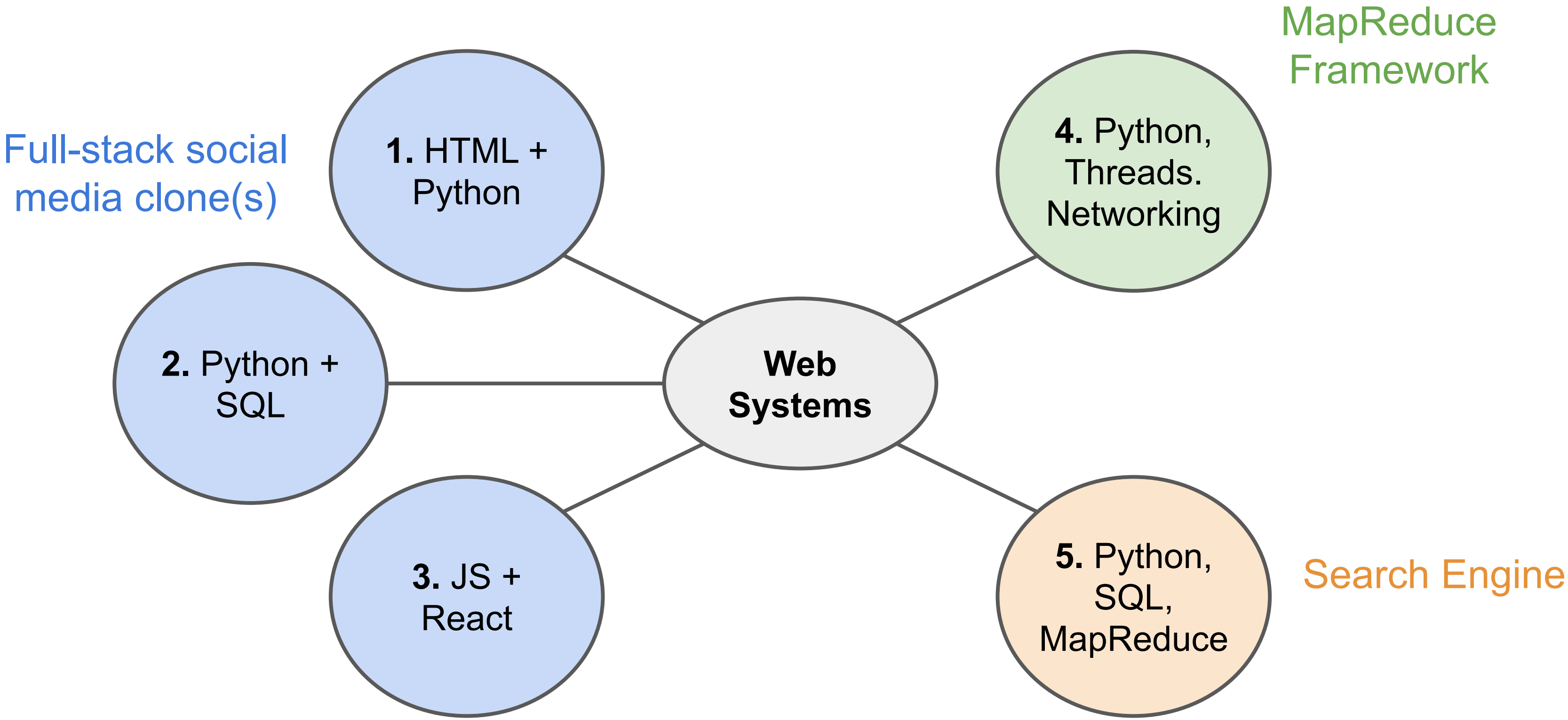
- Created 6 *specialist* RAG chatbots with varying scopes of course materials as external data
- Prompted each specialist chatbot and the *generalist* chatbot ChatGPT with sample student questions
- Analyzed hallucination and helpfulness between bot types using an evaluation by a team of expert instructors
    - Specialist vs generalist
    - Specialist (all-project / course-wide) vs specialist (one-project)

Course and Projects

- Upper-level Web Systems elective, 400+ enrolled each semester
  - Prerequisites: CS1, CS2, CS3, Discrete Math
- Most students are juniors and seniors

- 5 course programming projects covering full-stack web development, distributed systems, and search engines

Course and Projects

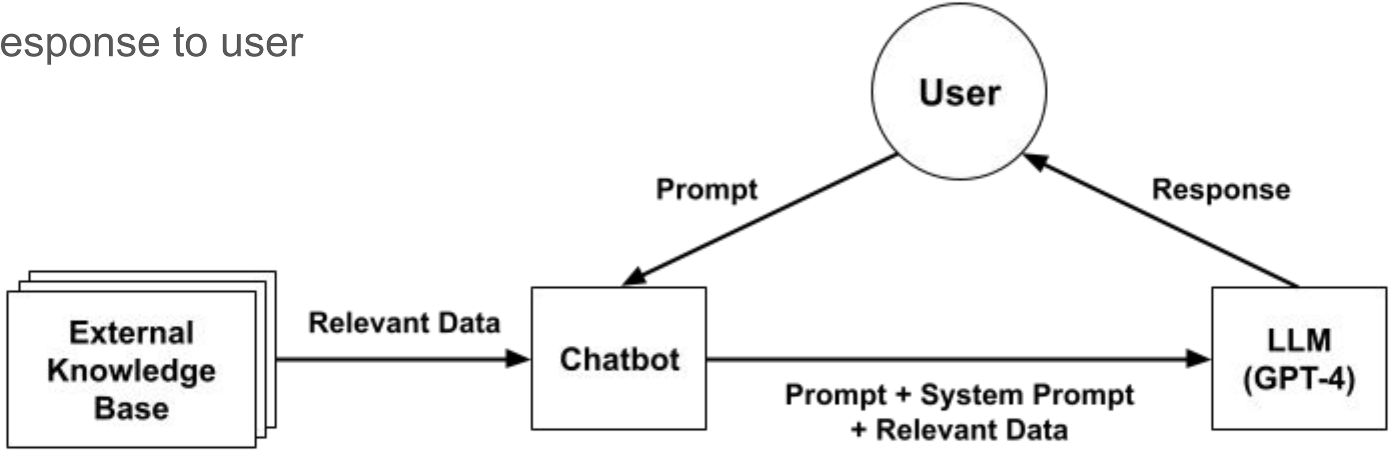


Specialist chatbots

All-project: Context covers all projects

One-project: Context covers one specific project

- 1 ***all-project*** and 5 ***one-project*** RAG chatbots
  - Searches for 4 most similar documents
  - Combines context documents with user prompt
  - Sends enhanced prompt to LLM (GPT-4)
  - Return response to user



External knowledge bases

- Collected course materials across all project topics:
    - 5 Project specifications
    - 15 Instructor-written tutorials
    - Slides and transcripts from 17 relevant lectures
    - Slides from 11 labs
    - Question-and-answer threads from relevant course forum posts over 3 semesters
- Documents were categorized by project and subsets were used to create knowledge bases for 5 one-project chatbots
- Knowledge base of all-project chatbot included all documents

Configuration

- Created unique teaching assistant "personas" by customizing system prompts with project details within scope

Imagine you are a helpful teaching assistant for a web development course. Your job is to help students with Project 1, an Instagram clone implemented with a templated static site generator in Python. The first part is hand-coding two web pages using static HTML. The second part is writing a Python program that renders static pages using the Jinja2 library. The third part is writing Jinja templates for all of the pages of the Instagram clone, following the requirements in our specification. 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. Keep responses as short as possible. {context} Question: {question} Helpful Answer:

Project 1 bot

Imagine you are a helpful teaching assistant for a web development course. The course has 5 projects: Project 1 is an Instagram clone implemented with a templated static site generator in Python. Project 2 is an Instagram clone implemented with server-side dynamic pages using Flask. Project 3 is an Instagram clone implemented with client-side dynamic pages, using React. Project 4 is a MapReduce framework distributed system in Python. Project 5 is a search engine with a MapReduce pipeline of programs using Python to generate an inverted index and a small server-side dynamic pages UI. 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. Keep responses as short as possible. {context} Question: {question} Helpful Answer:

All-project bot

Expert evaluation

- Set aside 340 (out of 2878) student questions from the course forum to use as sample prompts, labeled with project number
- For each question, the team of experts prompted 3 chatbots
    - All-project bot
    - One-project bot corresponding to the project the question was asked for
    - Generalist chatbot (ChatGPT Pro)
- Evaluated prompt and response quality using a coordinated rubric to compare chatbot performance and look for relationships between prompt and response quality

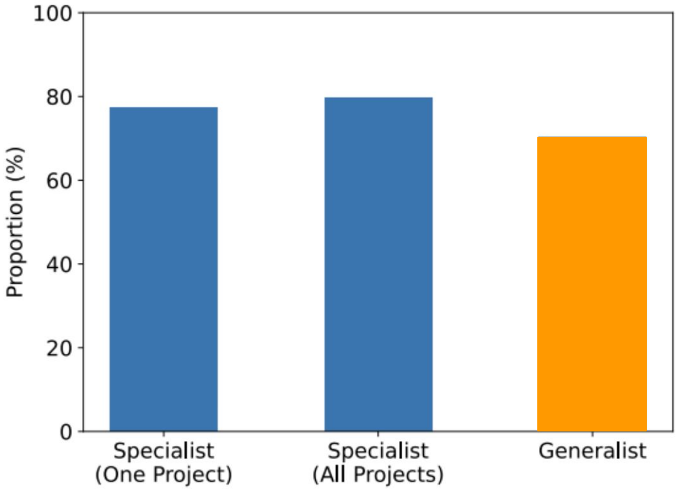
Expert evaluation

- **Prompt quality**
  - Clear = Can it be answered without follow up questions from a TA?
  - On-topic = Can a TA easily understand which project it is referencing?

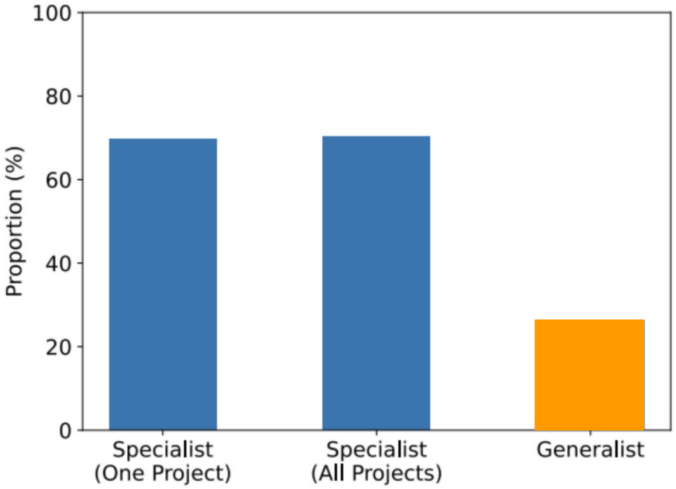
- **Response quality**
  - Correct = Is the answer correct (AKA not a hallucination)?
  - Helpful = Would the answer help students make progress?

Results – Specialists vs Generalist

- Specialist chatbots showed lower hallucination and higher helpfulness compared to the generalist



Correctness by bot type



Helpfulness by bot type

Results – Specialists vs Generalist

- Response length varied by chatbot type
    - Generalist produced the longest responses, both specialists produced responses with similar length that were shorter than the generalist

- **Example prompt**: For the human readable timestamps, should we have them updating in real time and without the page reloading? (or update but on page reload, or not update at all?)

Results – Specialists vs Generalist

All-project specialist: No, the timestamps do not need to update in real time without page reloading. They should update on page reload. Use ‘dayjs’ to format the timestamps when the component re-renders.

One-project specialist: You do not need to update timestamps in real-time. They should update based on the component re-rendering, for example, when a like or comment is added. Updating them on each component re-render will ensure the timestamp is reasonably accurate without the need for constant polling or real-time updates.

Generalist: For displaying human-readable timestamps on a webpage, the decision on whether to have them update in real time, update on page reload, or not update at all depends largely on the user experience you want to provide and the context in which the timestamps are used. Here are some considerations for each approach:

1. Updating in Real Time Without Page Reloading:

• Use Case: This is ideal for applications where the timing information is critical, such as in chat applications, live event tracking, or monitoring dashboards.

• Benefits: It keeps the displayed time accurate without requiring any interaction from the user, enhancing the dynamism and interactivity of the application.

[... shortened for brevity]

Specialist chatbots generated concise and straightforward answers to student questions, while generalist produced much longer, generic responses with non-specific advice

Results – Specialists vs Generalist

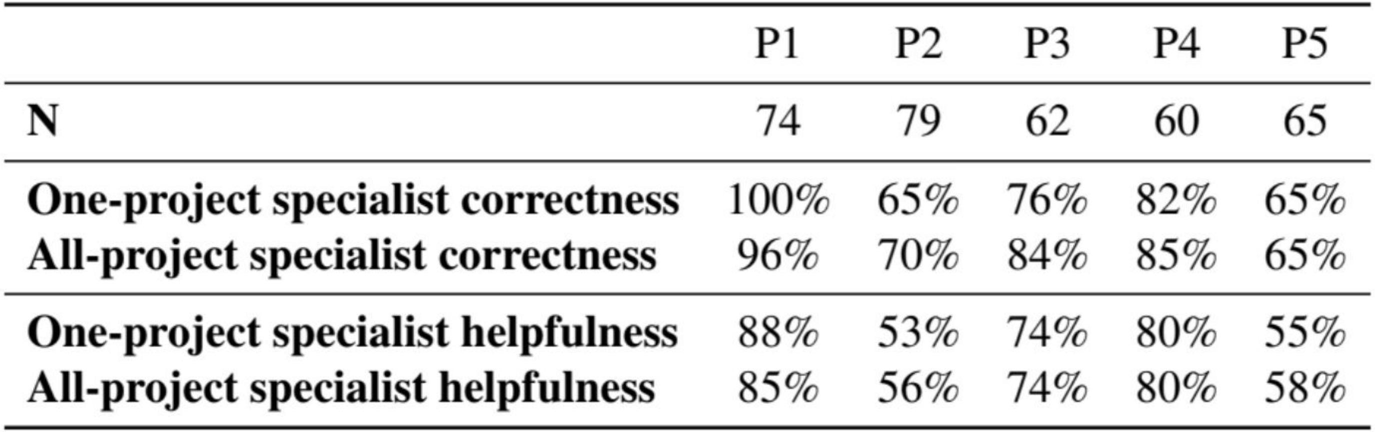
- Why do RAG chatbots lead to less hallucinations and more helpful responses to domain-specific prompts?
    - **High-quality retrieval system + high-quality documents returned and summarized = high-quality response**

- Correct and concise responses were considered more helpful by team of experts, even if lengthy generalist response contained partially correct information

Specialist chatbots were more correct and more helpful than generalist chatbots for course projects in this study

Results – Specialist scope comparison

- Per project, all-project and one-project specialist chatbots had about the same rates of correctness and helpfulness
  - Some projects had worse performance than others (likely due to project topic complexity and question quality)



	P1	P2	P3	P4	P5
N	74	79	62	60	65
One-project specialist correctness	100%	65%	76%	82%	65%
All-project specialist correctness	96%	70%	84%	85%	65%
One-project specialist helpfulness	88%	53%	74%	80%	55%
All-project specialist helpfulness	85%	56%	74%	80%	58%

Results – Specialist scope comparison

- All-project chatbot retrieved on-topic documents from its knowledge base despite its larger size with more distracting documents
    - Documents in combined knowledge base were fairly disjoint since project topics were distinct
    - Retrieval system did not have extra difficulty finding relevant documents compared to one-project chatbots with limited knowledge bases
- Relevant documents returned were similar to documents returned by one-project bots, resulting in similar outputs for a given prompt

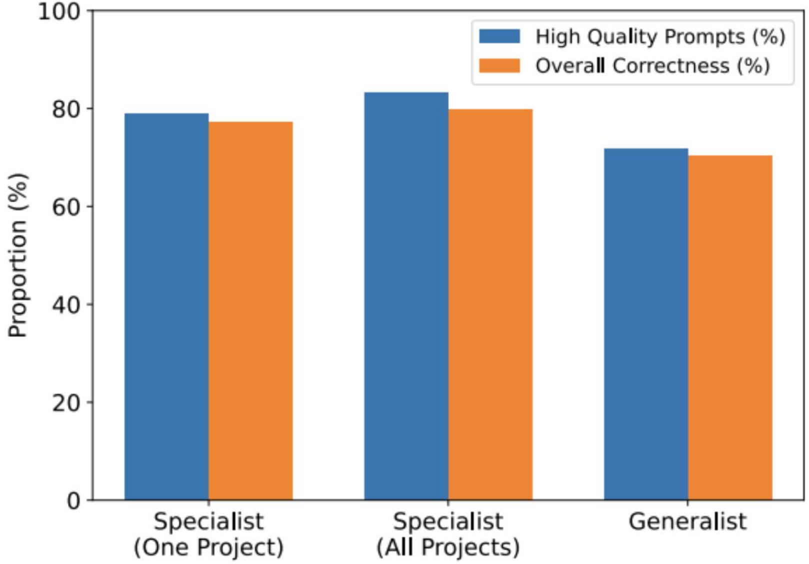
The scope of a specialist chatbot does not appear to affect its rates of hallucination or helpfulness

Results – Impact of prompt quality

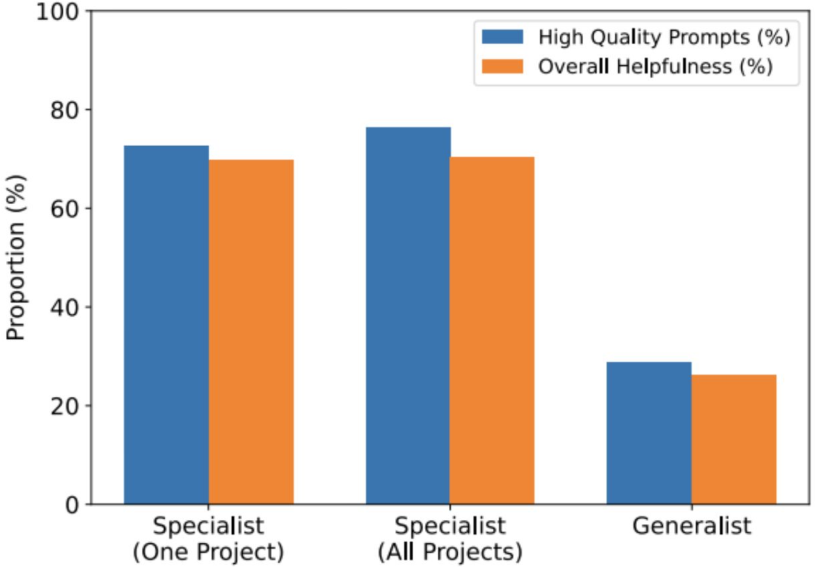
- Large majority of sample prompts were high-quality
  - 84.4% clear and 90.9% on-topic
- Found a statistically significant association between prompt quality and both the correctness and helpfulness of responses
- All types of chatbot performed marginally better on high-quality prompts, but not enough to justify extra efforts to exclusively use high-quality prompts in experiments

Prompt quality had a minimal impact on response quality, but high-quality prompts are ideal when available

Results – Impact of prompt quality



Correctness of responses on high-quality prompts per bot type. All bots had slightly higher rates on high-quality prompts than their overall correctness.



Helpfulness of responses on high-quality prompts per bot type. All bots had slightly higher rates on high-quality prompts than their overall helpfulness.

Limitations

- Scope of the study is limited to one upper-level course
- Any supplemental images in original prompts were not provided to bots in experiment
- Prompts from course forum were intended for a human audience (TAs); students may phrase questions differently for a chatbot

Conclusions

- We found:
  - Specialist chatbots hallucinated less and provided more helpful answers than the generalist chatbot (RQ1, RQ2)
  - Scope of external data for knowledge base did not significantly impact the quality of responses (RQ3)
  - Prompt quality also had minimal impact on response quality
- Our results provide insights for instructors considering one (or multiple) RAG chatbots as a course resource over a general-purpose chatbot
  - One course-wide bot is likely to be as effective as multiple project-scoped bots