Specialist vs. Generalist Generative AI for Student Learning

Melina O'Dell and Andrew DeOrio
Department of Electrical Engineering and Computer Science
University of Michigan

1 Abstract

Generative AI (GenAI) chatbots have quickly become a popular and powerful tool. To provide students with course-specific help, educators may build custom retrieval-augmented generation (RAG) chatbots with external knowledge bases comprised of course resources. Do these "specialist" chatbots provide advantages over general-purpose chatbots? How does the scope of the external data affect the helpfulness of the responses?

We created a suite of specialist custom chatbots: one with a wide scope encompassing all programming projects in a course and five with limited scope, each specialized to one project. Each chatbot was given an external knowledge base containing the project specification, tutorials, lecture content, lab materials, and past course forum posts to reference. The course is a high-enrollment Web Systems elective at a large public research university.

We gathered a subset of sample student questions from the course forum to use as prompts. An expert team of instructors evaluated each bot's response for accuracy (hallucination) and helpfulness. We used ChatGPT Pro as a baseline "generalist" chatbot.

Overall, the specialist bots hallucinated less and were considered more helpful for student questions than the generalist bot. The best specialist bot was correct 80% of the time and helpful 70% of the time. The generalist bot was correct 70% of the time and helpful only 26% of the time. There were minimal performance differences between the specialist bots with varying scopes.

Our experience can guide educators using generative AI. First, a custom RAG chatbot is more helpful than a general-purpose chatbot. Second, a single chatbot with a course-wide scope has a similar effectiveness to multiple narrower-scope chatbots.

2 Introduction and Related Work

As generative AI tools grow in popularity and ability, students and instructors are exploring their role in education. Chatbots like OpenAI's ChatGPT [1] are commonly used by students to assist with their coursework, such as writing and programming tasks [2]. Educators are also trying to integrate chatbots into their courses as an additional resource, as they are known to enhance academic performance when utilized properly [3], [4].

Generative AI tools struggle with hallucinations, or incorrect or misleading responses [5]. These false chatbot responses could compromise a learning environment, as students may be unaware and mistakenly trust them. In an attempt to reduce frequent hallucination, retrieval-augmented generative AI (RAG) systems were introduced. These systems use an external knowledge base to provide additional context beyond the user-provided prompt to the LLM and have demonstrated reduced hallucination in many cases [6]. OpenAI released their own RAG chatbots, called GPTs [7], to provide a gateway to customized chatbots for domain-specific tasks.

A previous study on introducing a custom chatbot as a course resource to students found that it performed significantly better than a general-purpose chatbot (ChatGPT) when providing course-specific help on one assignment [8]. However, university courses typically have more than one assignment, raising questions about the optimal scope of external data provided to a custom course help chatbot. Would one chatbot referencing all materials from a course perform well on any assignment, or suffer from more hallucination because it is less focused? Would an assignment-specific chatbot perform better because it has a narrower scope for that assignment, or is maintaining several chatbots for a course overdoing it? We aim to explore these questions with our experiments.

2.1 Contributions

The goal of this study is to analyze differences in chatbot performance with different external data scopes. We will also compare "specialist" RAG chatbots with varying scopes with a "generalist" state-of-the-art, general-purpose chatbot. In the context of student learning augmented by a generative AI course resource, our research questions are:

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?

3 Methods

First, we introduce the course context in which our study took place. Then, we will describe the generative AI chatbots used in our experiments. We used two kinds of specialist chatbots: one with a narrow scope focused on a single project and another a wider scope covering all the projects. For the generalist bot, we used ChatGPT Pro. Finally, we outline the methods for prompt and response evaluation by a team of expert instructors.

3.1 Context

The course is a high-enrollment Web Systems elective at a large public research university. Prerequisite courses include data structures and algorithms, programming (CS2), and discrete math. Most students in the course were juniors and seniors.

The project topics in the course cover full-stack web development, distributed systems, and search engines. The first three projects involve building a full-stack social media clone. The next project is a distributed system where students implement a MapReduce framework using basic parallel

and networking programming techniques in Python. The final project is a search engine, where students implement a a small web application and a sequence of MapReduce programs to build an inverted index.

3.2 Generalist chatbot

We used OpenAI's ChatGPT Pro as the generalist chatbot in our study. It used the GPT-4 model [9] and did not have an external knowledge base to reference for additional context.

3.3 Specialist chatbots

We created six custom retrieval-augmented generative AI (RAG) chatbots for our study, each with a different external knowledge base. Five bots focused on one project each, while the sixth had data on all projects. At query time, the chatbot searches for the most relevant documents to the user prompt, then combines them with the original prompt and a system prompt. This enhanced prompt is sent to a Large Language Model (LLM) and the response is returned to the user.

3.3.1 Chatbot architecture

The specialist chatbots are based on OpenAI's GPT-4. When a user prompts a chatbot, it performs a similarity search over its external knowledge base using a vector model, created using an embedding model provided by OpenAI for use with GPT-4. It then selects up to four of the closest documents and combines them with the user's prompt as context. This enhanced prompt is sent to an instance of GPT-4 to generate a response.

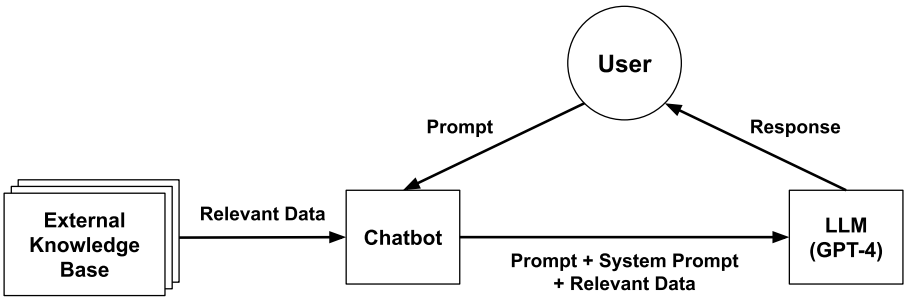


Figure 1: Specialist RAG chatbot architecture. The bot searches for relevant documents and sends them to GPT-4 along with the user's prompt and a system prompt.

3.3.2 External knowledge bases

Each RAG chatbot was given a dedicated collection of external data containing course materials relevant to the projects in our study. These documents are searched at query time to enhance the user prompt by providing additional context about the course projects to GPT-4. Each specialist chatbot returns a list with the four documents it chose along with its response.

We collected 2575 documents across all five projects to use as external data, consisting of five different types of course material made available to all students:

- 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

Each thread in the course forum posts included a student question, instructor answer, student answer, any follow up discussions, and a label corresponding to one of the five projects.

We collected 2954 student questions from the course forum. After filtering invalid questions, we had 2878 questions remaining. We set aside 340 of these to use as sample prompts for evaluation (using the standard formula for calculating sample size with 5% margin of error, 95% confidence level, and population proportion of 50%), leaving 2538 to use as external data.

The one-project specialist chatbots were provided with the subset of the materials related to the project of their focus, while the all-project specialist chatbot was provided with the entire collection, containing all of the documents for the course. Table 1 shows the distribution of external data per chatbot.

	All-project specialist	One-project specialist				
		P1	P2	P3	P4	P5
Project Specs	5	1	1	1	1	1
Tutorials	15	3	3	5	4	6
Lectures	17	2	3	4	4	4
Labs	11	2	2	3	2	2
Q/A Threads	2538	647	535	474	495	387

Table 1: External data provided to each of the specialist chatbots. The one-project specialist chatbot used data from one project while the all-project specialist chatbot used the combined data. Some materials (such as tutorials) spanned multiple projects.

3.3.3 System prompts

The system prompt defines the way a chatbot responds to users by creating a "persona." We configured our chatbots to act as a helpful teaching assistant for a web development course.

We customized the system prompt for each RAG chatbot. Each of the specialist chatbots was given a prompt with specifics related to the project of their focus (or a summarized description of all five projects in the case of the all-project specialist chatbot, to ensure it had comparable context on each project). Figure 2 shows the system prompts for the all-project specialist chatbot and a one-project specialist chatbot, where *context* and *question* are replaced with the additional context from the external knowledge base and the user's original prompt.

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:

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:

Figure 2: The system prompts for the all-project (left) and one-project (right) specialist bots.

3.4 Expert evaluation

We evaluated the performance of our bot suite with a team of expert instructors who are familiar with the course material. We selected a subset of 340 sample student questions from the course forum to use as prompts for the chatbots.

The team of experts used a coordinated rubric (Sections 3.4.1 and 3.4.2) to evaluate prompts and responses. For each student question in our dataset, one instructor prompted the all-project specialist chatbot, the one-project specialist chatbot corresponding to the project the prompt was intended for, as well as the generalist chatbot ChatGPT Pro. The expert then documented four observations: Prompt quality, response quality, the top four relevant documents returned with the response, and which of the chatbots they believed performed the "best" qualitatively. We used this data to determine the accuracy and helpfulness for each bot and measure hallucination.

3.4.1 Prompt evaluation

We trained our expert instructors to evaluate the quality of the prompts. These results are used to look for relationships between the quality of the prompts and the chatbot responses.

We defined prompt quality as a combination of "clear" and "on-topic." A clear question contained enough information in the prompt to understand what the student was confused about without the need to ask followup questions to return a correct and helpful response. An on-topic question contained enough information specific to the project for an instructor to determine which of the five projects the question was about. A "high-quality" prompt was both clear and on-topic.

High-quality prompt example: "Passing Locally but not Autograder: Hello, my partner and I are passing the index file locally but not through the autograder and we're not sure what the issue is. Any help is appreciated! Thanks"

Low-quality prompt example: "Existence of templates directory Do we need to throw an error if the templates directory does not exist? Or can we assume that it will always be present?"

3.4.2 Response evaluation

We also trained our expert instructors to evaluate the quality of the responses using their knowledge of the course. These results are used to compare the performance between the three types of chatbots.

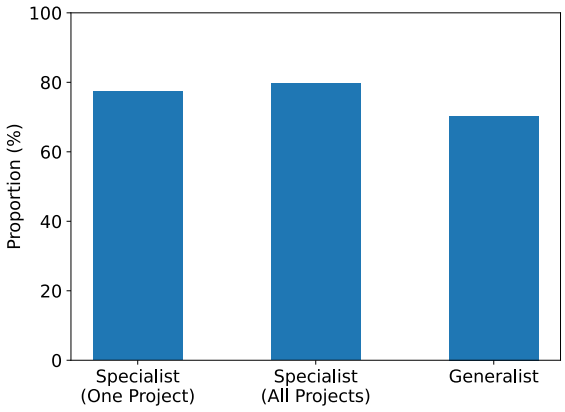
We defined response quality as a combination of "correctness" and "helpfulness." A correct response is one that is correct considering the student's prompt in isolation. A helpful response is one that would help students make progress on the project and would be considered reasonable advice given by course staff. Example responses are shown in Section 4.1.

4 Results

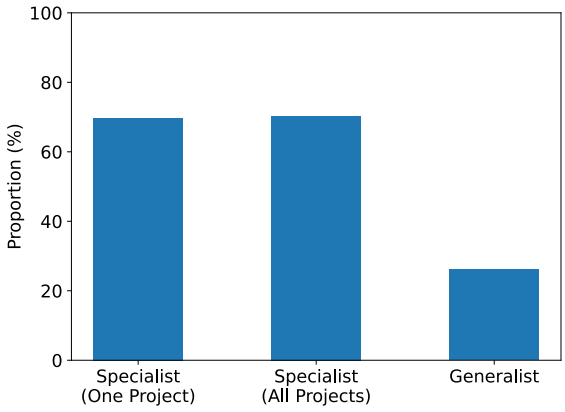
First, we provide general statistics for the accuracy (hallucination) and helpfulness of each chatbot. Then, we look at the impacts of prompt quality and the document retrieval system on response quality. Finally, we compare the performance of the specialist chatbots.

4.1 Hallucination and helpfulness

Overall, our RAG-based specialist chatbots demonstrated low rates of hallucination and high rates of helpfulness. The all-project specialist bot answered 80% of sample questions correctly and 70% were helpful to students. Across our one-project specialist bots, 77% of questions were answered correctly, and 70% were helpful to students. The generalist bot, ChatGPT, did not perform as well, answering 70% of questions correctly, but only 26% of responses were considered helpful (Figures 3a and 3b).



(a) Correctness of chatbot responses by bot type. The specialist chatbots performed similarly to each other and slightly better than the generalist chatbot.



(b) Helpfulness of chatbot responses by bot type. The specialist chatbots performed similarly to each other and significantly better than the generalist chatbot.

Figure 3: Comparison of overall chatbot correctness and helpfulness.

The generalist chatbot tended to produce the longest responses, suggesting a broad range of vague options to answer the student prompt. The specialist chatbots generated responses that were much more concise, straightforward, and confident regarding course topics. There was no distinct qualitative difference between the performance of the all-project and one-project specialist chatbots. Figure 4 demonstrates these observations with example responses from each bot to a student prompt about updating timestamps on a webpage.

> **Prompt:** timestamp updating: 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?)
>
> **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]*
>
> Each method has its own advantages, and the choice should align with the needs of your application and its users. If your content benefits from showing how current it is or if it's critical for users to know the exact moment data was last updated, real-time updates can be very beneficial. On the other hand, if the timestamp merely serves as a historical reference, static timestamps might be entirely adequate.

Figure 4: Example responses from each bot for a Project 3 student prompt. The specialist bots provided concise, actionable responses. The generalist bot was less helpful, instead providing a long list of suggestions less specific to the project.

We performed chi-square tests to evaluate the association between the type of bot (one-project specialist, all-project specialist, generalist) and the correctness and helpfulness of its responses (Table 2). We found a statistically significant difference in correctness and helpfulness between our specialist chatbots and ChatGPT, but not among our specialist chatbots.

Our team of experts labeled the responses of the one-project specialist bots as "best" most often, for over 58% of prompts, followed closely by the all-project specialist bot with 56% (Table 3).

Bot types	p-value	χ^2 statistic	df
Specialists vs. generalist and correctness	0.012	8.89	2
One-project specialist vs. all-project specialist and correctness	0.512	0.43	1
Specialists vs. generalist and helpfulness	5.483e-39	176.20	2
One-project specialist vs. all-project specialist and helpfulness	0.933	0.01	1

Table 2: Chi-square test results analyzing the association between the type of bot and the correctness and helpfulness of responses. We observed statistically significant associations between our custom chatbots and the generalist ChatGPT, but not between our custom chatbots.

Bot type	% "Best Bot" responses
All-project specialist	56.47%
One-project specialist	58.82%
Generalist	8.53%

Table 3: Ranking of "Best Bot" after expert review. The one-project specialist bots were rated the best the most times, closely followed by the all-project specialist bot. The generalist bot was rarely the best.

Many of the responses from the specialist chatbots were meaningfully equivalent and better than the generalist bot responses, so our experts labeled them both as tied for the "best."

4.2 Impact of prompt quality

Out of 340 prompts overall, 84.4% were clear and 90.9% were on-topic after expert review. Prompt quality varied by project (Table 4).

Project	% Clear	% On-topic
P1	81%	100%
P2	77%	85%
P3	90%	82%
P4	88%	95%
P5	88%	92%

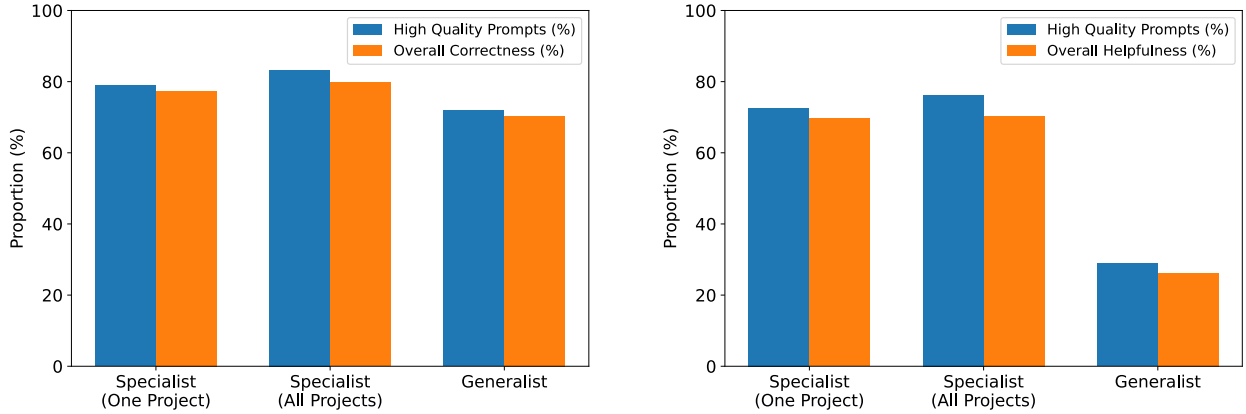
Table 4: Percentage of clear and on-topic prompts per project.

We performed chi-square tests to see if prompt quality was associated with response quality (correctness and/or helpfulness) (Table 5). We found that there is a statistically significant association between prompt quality and both the correctness and helpfulness of responses.

Figure 5a compares the percentage of correct responses for each type of chatbot on high-quality prompts with the overall rate of correctness. All chatbots performed slightly better on high-quality prompts. Similarly, Figure 5b shows the same comparison for the rate of helpfulness and also displays a slight increase for high-quality prompts.

Bot types	p-value	χ^2 statistic	df
Clear and correct	0.012	6.26	1
Clear and helpful	5.855e-05	16.15	1
On-topic and correct	1.053e-05	19.41	1
On-topic and helpful	0.0002	14.09	1

Table 5: Chi-square test results analyzing the association between prompt quality metrics (clear and on-topic) and the correctness and helpfulness of responses. We observed statistically significant associations between prompt quality and both correctness and helpfulness.



(a) Correctness of bot responses on high-quality prompts. All bot types had a slightly higher rate of correctness on high-quality prompts than their overall correctness.

(b) Helpfulness of bot responses on high-quality prompts. All bot types had a slightly higher rate of helpfulness on high-quality prompts than their overall helpfulness.

Figure 5: Comparison of chatbot correctness and helpfulness on high-quality prompts.

4.3 Comparison of specialist bots

We observed differences in performance for the specialist bots on different projects. The specialist bots had the highest rates of correctness and helpfulness on Project 1, while Projects 2 and 5 had the lowest (Table 6). When comparing the performance of the one-project specialist bots with the all-project specialist bot within the same project, they had about the same rates of correctness and helpfulness.

4.3.1 Impact of retrieved documents

Most responses from the specialist bots referenced at least one question-and-answer thread from the course forums to produce helpful responses. However, per project, the project specification was almost always the most frequently referenced document.

We calculated the intersection between documents returned by the one-project and all-project specialist chatbots per prompt to analyze the performance of the retrieval system. On average, 41.25% of documents returned by the bots overlapped.

	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%

Table 6: Correctness and helpfulness between specialist bots per project. The performance of the all-project specialist bot is in line with the performance of the one-project specialist bots. The Project 1 specialist bot had the highest rates of correctness and helpfulness.

5 Discussion

Overall, the specialist chatbots were less likely to hallucinate and more helpful than the generalist chatbot (RQ1 and RQ2). Neither the scope of the external data provided to the specialist chatbots nor the prompt quality substantially impacted performance (RQ3).

5.1 Specialist chatbots were more correct and helpful than generalist chatbots

The specialist chatbots in our study provided more correct and more helpful responses than the generalist chatbot. This suggests that RAG chatbots, with an external knowledge base to enhance responses, lead to less hallucinations and more helpful responses to the domain-specific prompts in our study.

The specialist chatbots excelled at summarizing the information provided by their retrieval systems. Providing high-quality, focused documents in our knowledge bases allowed the retrieval system to target the most relevant sources and produce a highly relevant response. It is likely that a RAG system with a poor retrieval system would tend to return irrelevant documents, producing irrelevant and unhelpful responses.

The length of a bot response may impact the likelihood of its correctness. The team of experts found that sometimes generalist bot responses were only considered correct because they were very long and contained many possible answers (including the correct one). However, these responses were still considered unhelpful because they included extra information that could be misleading to students. By contrast, the specialist chatbots answered correctly and concisely for most prompts, which is considered more helpful.

5.2 Specialist chatbots responded similarly despite varying external data scopes

The specialist chatbots tended to respond similarly for a given prompt. Our experts reported qualitative similarities between the one-project and all-project specialist responses. There were also similar rates of correctness and helpfulness between the one-project specialist and all-project specialist for a given project. These results suggest that the scope of a specialist chatbot did not affect its rates of hallucination or helpfulness.

We found that the all-project chatbot tended to retrieve on-topic documents despite its larger knowledge base, indicating that its retrieval system was effective. The one-project bots had a less complex retrieval task than the all-project bot because their knowledge bases were smaller and free of many distracting documents. One contributor to the effective retrieval system of the all-project bot was limited overlap between the knowledge base documents. The documents in the combined external knowledge base were fairly disjoint in content because the topics for each of the projects were distinct. This meant that the retrieval system was more likely to find and return relevant documents. When correct and relevant documents are returned, similar summarizations of these documents are created for the response, leading to the similar output of the specialist bots.

5.3 Prompt quality had minimal impact on response quality

Each chatbot performed marginally better on high-quality prompts than low-quality prompts. There also were no large differences on high-quality prompts based on the type of chatbot (specialist or generalist). This suggests that prompt quality did not impact response quality enough to justify specific efforts to use high-quality prompts with the chatbots in this study.

5.4 Limitations

The prompts used in our study came from student questions on the course forum. These questions were intended for a human audience (course staff). Students may phrase questions differently when knowingly interacting with chatbots than in our sample. Additionally, many prompts originally included images that were not provided to the chatbots in our experiments, but would be available to staff on the forum.

Chatbot outputs are non-deterministic. The qualitative English output that our experts evaluated could have been different with repeated trials or another variation of prompt wording.

Bot performance evaluations were subjective and based on expert instructor experiences. Course staff can draw on prior experience doing the projects themselves to identify common problems and help students, whereas the bots can only use the information provided in the prompt.

Finally, our results are from one upper-level computer science course. A lower-level class with more of its content documented on the Internet may see the generalist bot perform better.

6 Conclusions and future work

We created six specialist RAG chatbots with varying scopes to assist students with projects in an upper-level Web Systems course. Each chatbot was provided with a different external knowledge base, focusing on either one project or all projects in the course, to use as additional context when answering student questions. We compared specialist and generalist (ChatGPT) chatbot performance using an evaluation of the prompts and responses by a team of expert course instructors.

We found that the specialist chatbots hallucinated less and provided more helpful answers than the generalist chatbot when given student prompts. The scope of the external data provided to

each specialist bot did not appear to affect the quality of responses. The quality of the prompt also did not seem to impact response quality. We did not observe performance gains significant enough to justify the additional effort of creating tighter-scoped chatbots over one chatbot with a course-wide scope.

The implications of our study for instructors include considering the use of a custom, RAG-based chatbot over a general-purpose chatbot as a course resource. One bot that covers multiple assignments, or even the whole course, is likely to be as effective as multiple narrower-scope bots.

Future work could perform our experiment on projects from a lower-level computer science course to analyze if generalist chatbots see performance improvements on highly-documented concepts. It would also be useful to examine whether having projects and external data with more overlap (ours were fairly disjoint) would amplify the performance differences between the specialist bots with different scopes.

7 Acknowledgments

We thank Yutong (Ellen) Ai and Akanksha Girdhar for their contributions to the expert evaluation.

References

[1] OpenAI, "ChatGPT." https://openai.com/index/chatgpt/, 2024.

[2] C. Baek, T. Tate, and M. Warschauer, ""ChatGPT seems too good to be true": College students' use and perceptions of generative AI," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100294, 2024.

[3] Y. Xue, H. Chen, G. R. Bai, R. Tairas, and Y. Huang, "Does ChatGPT Help With Introductory Programming? An Experiment of Students Using ChatGPT in CS1," ICSE-SEET '24, (New York, NY, USA), p. 331–341, Association for Computing Machinery, 2024.

[4] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu, "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies," *Computers & Education*, vol. 227, p. 105224, 2025.

[5] H. Alkaissi and S. Mcfarlane, "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing," *Cureus*, vol. 15, 02 2023.

[6] P. Feldman, J. R. Foulds, and S. Pan, "RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots," 2024.

[7] OpenAI, "GPTs." https://openai.com/index/introducing-gpts/, 2023.

[8] Y. Ai, M. Baveja, A. Girdhar, M. O'Dell, and A. DeOrio, "A Custom Generative AI Chatbot as a Course Resource," 2024.

[9] OpenAI, "GPT-4." https://openai.com/index/gpt-4/, 2024.