References (delete or hide for presentation)

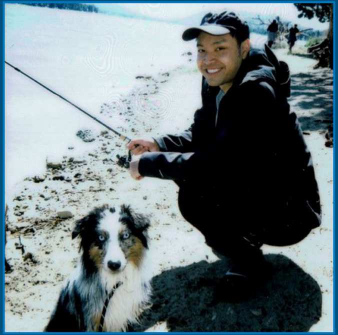
- Rubric with tasks - link
- Team project plan - link
- Roadmap - link

MIDSearch

A RAG based multiplatform search tool for the MIDS program

Devin Suy, Nadia Tantsyura, Randy Louie, Robert Greer, Thomas Lai

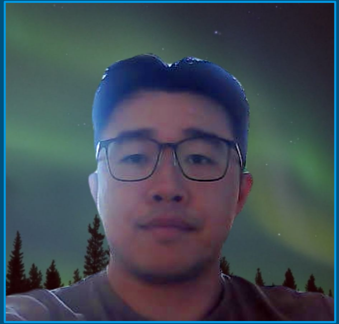
Team



Devin Suy

Nadia Tantsyura

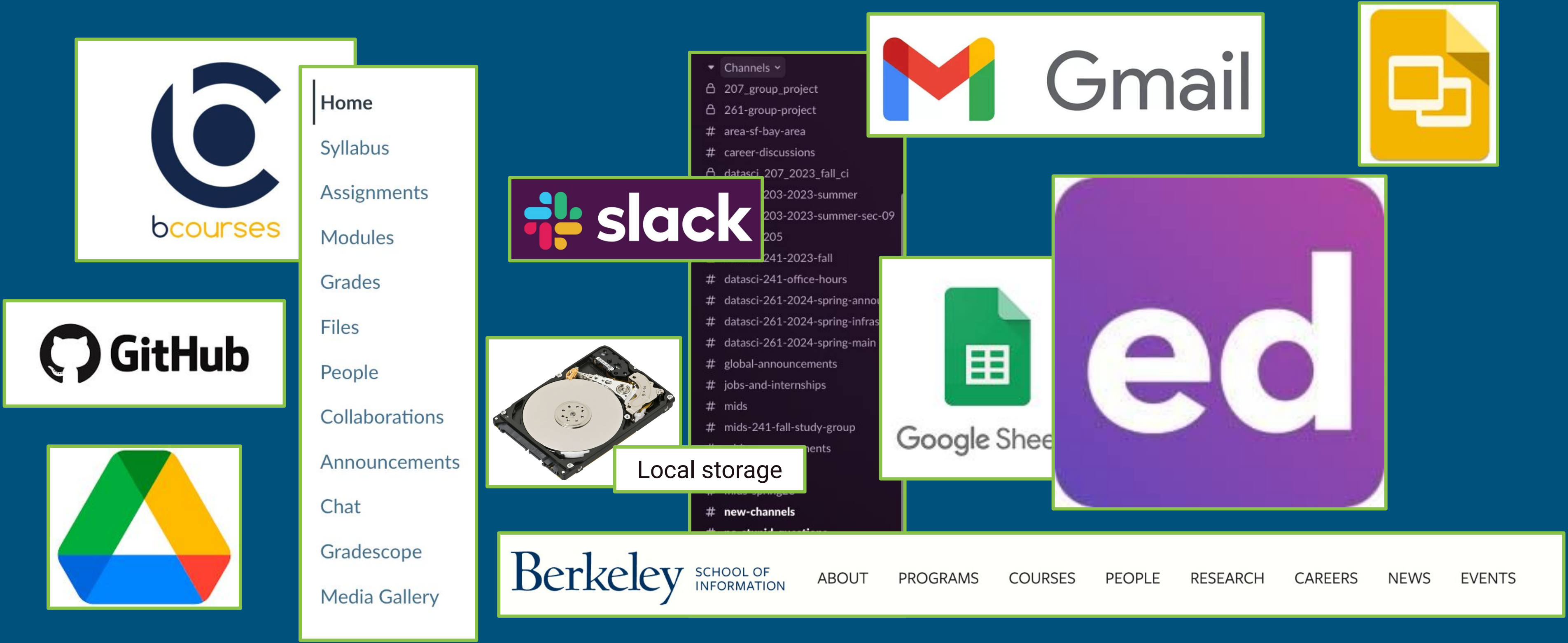
Thomas Lai



Randy Louie

Robert Greer

Problem & Motivation



Home
Syllabus
Assignments
Modules
Grades
Files
People
Collaborations
Announcements
Chat
Gradescope
Media Gallery

Channels
207_group_project
261-group-project
area-sf-bay-area
career-discussions
datasci_207_2023_fall_ci
203-2023-summer
203-2023-summer-sec-09
205
241-2023-fall
datasci-241-office-hours
datasci-261-2024-spring-annou
datasci-261-2024-spring-infras
datasci-261-2024-spring-main
global-announcements
jobs-and-internships
mids
mids-241-fall-study-group
ents
new-channels

Berkeley SCHOOL OF INFORMATION
ABOUT PROGRAMS COURSES PEOPLE RESEARCH CAREERS NEWS EVENTS

Market research

Knowledge bases / Chat apps
(non-cohesive)

slack

ATLASSIAN

OneDrive

GitHub

G Suite

Discord

Local storage

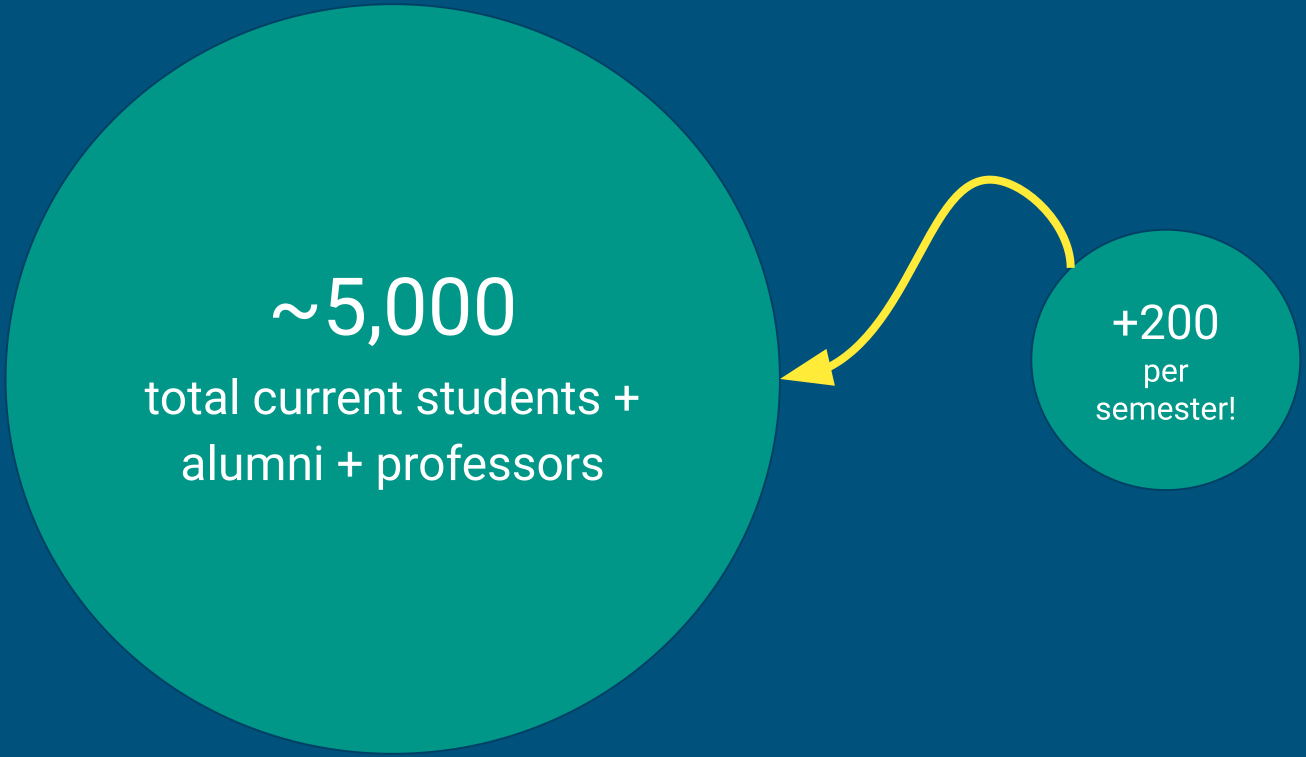
Cohesive-ish apps

Microsoft Copilot

Recall function
(releasing 6/18/24!)

glue.

Market size



~5,000

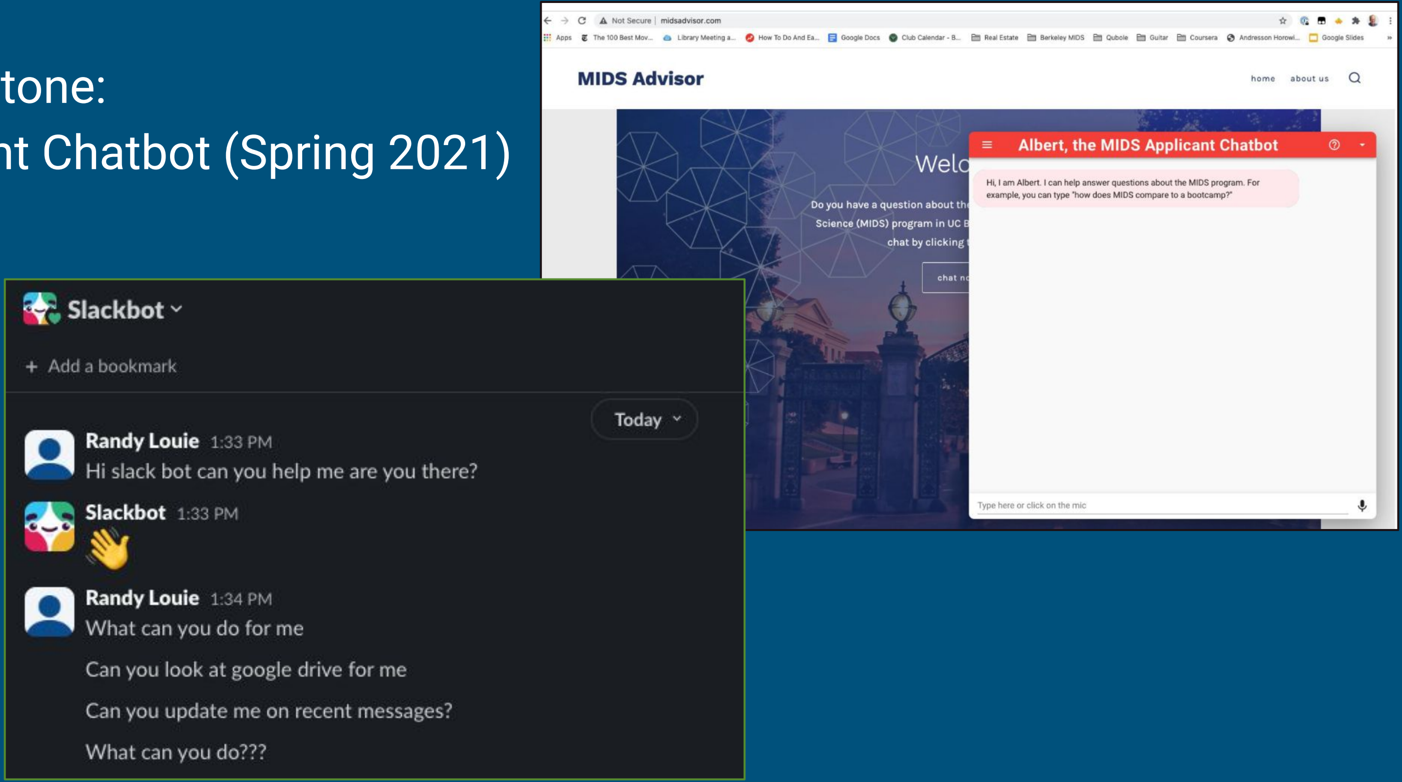
total current students + alumni + professors

+200

per semester!

What has been tried before

- Previous capstone:
  MIDS Applicant Chatbot (Spring 2021)
- Slack bot 👎



Not Secure | midsadvisor.com
MIDS Advisor
home
about us
Welc
Do you have a question about th
Science (MIDS) program in UC B
chat by clicking
chat no
Albert, the MIDS Applicant Chatbot
Type here or click on the mic
Slackbot
+ Add a bookmark
Today

Hi, I am Albert. I can help answer questions about the MIDS program. For example, you can type "how does MIDS compare to a bootcamp?"

Randy Louie 1:33 PM
Hi slack bot can you help me are you there?

Slackbot 1:33 PM
👋

Randy Louie 1:34 PM
What can you do for me

Can you look at google drive for me

Can you update me on recent messages?

What can you do???

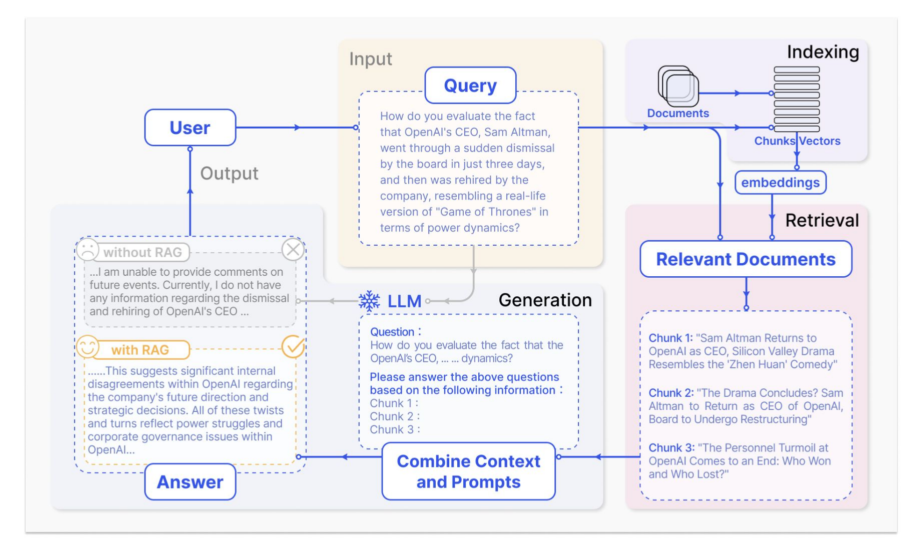
Our Solution and Improvements

RAG based system : Retrieval-Augmented generation is the process of optimizing the output of a large/small language model, so it references an authoritative knowledge base outside of its training data sources before generating a response.

Small Language Model : similar to Large language Model but smaller

User type and customers : Students and Instructors.

Goal is to Reduce implementation and training costs, improve reliability, improve security



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

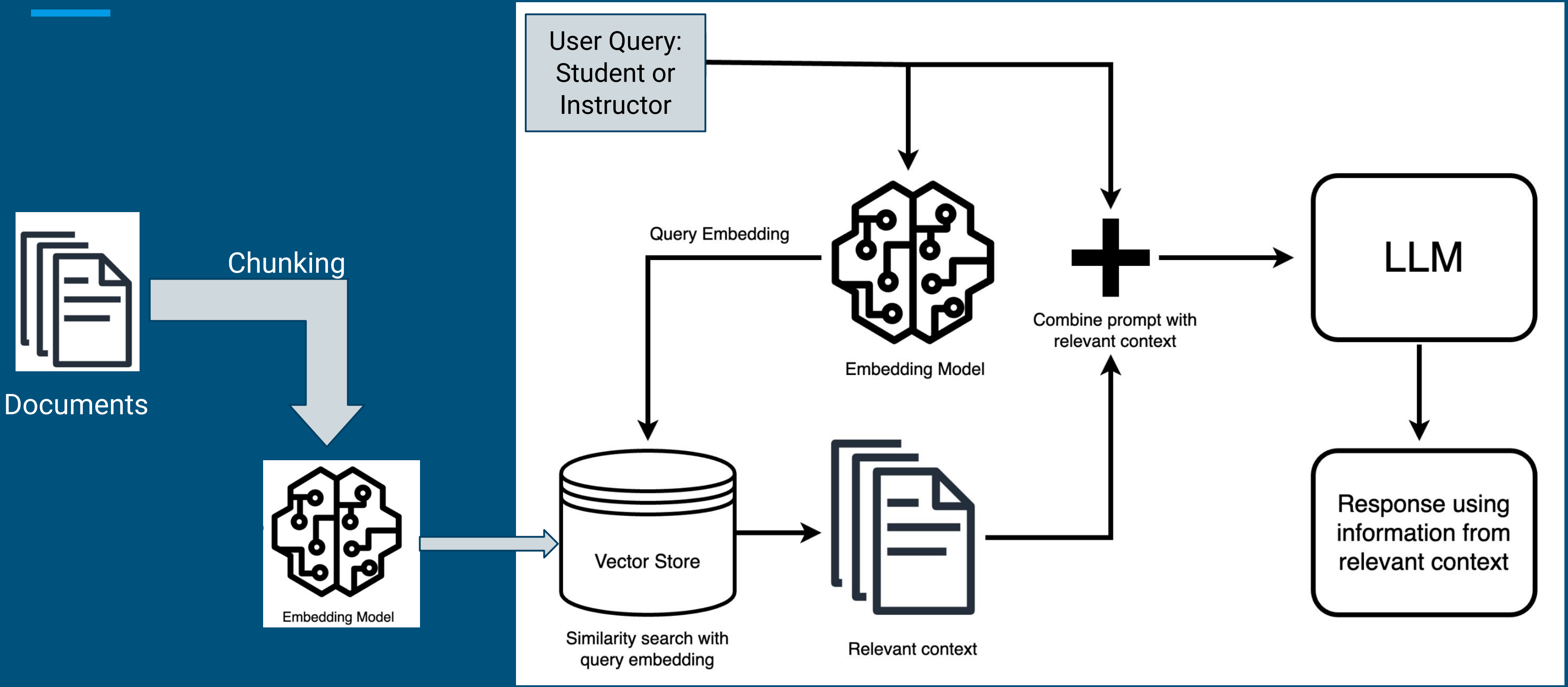
Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Deep Dive into RAG components



User Query: Student or Instructor

Documents

Chunking

Embedding Model

Query Embedding

Embedding Model

Combine prompt with relevant context

LLM

Vector Store

Similarity search with query embedding

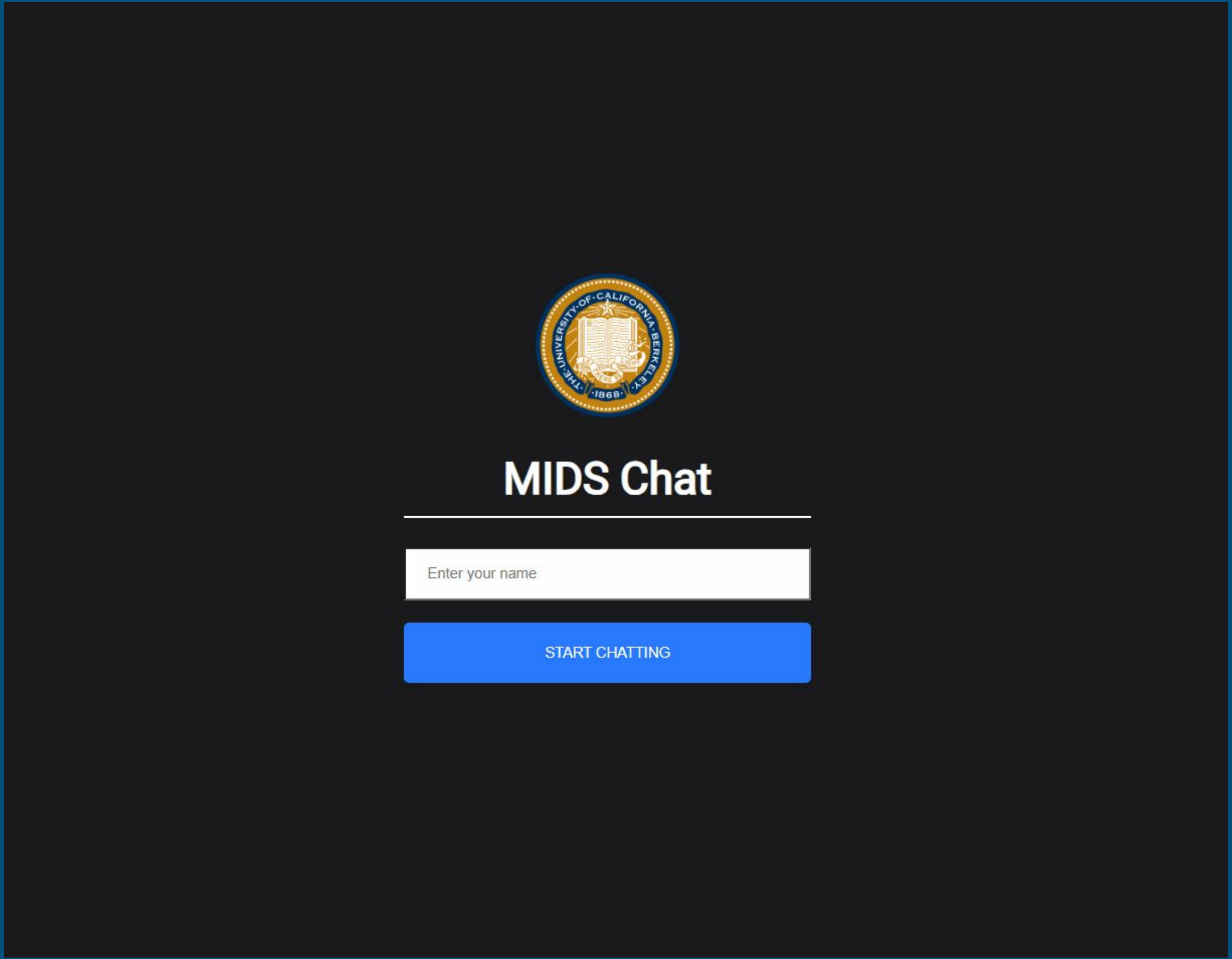
Relevant context

Response using information from relevant context

Roadmap - Yes, it’s week 5 already

	Week 7	Week 10	Week 12	Week 14
Language Model	Baseline LLM selected SLM Settings Default settings implemented	LoRA User Based Prompting Model Compression Techniques	Prompt Tuning Hyperparameter Tuning	API Chaining
Retrieval	Vector Database Establish Defaults	Search Strategies Reranker	Dynamic Chunking Hybrid Vector Database	Establish Protected Documents
Data	Collection EDA Survey Users	Generate User Specific Responses PDF Processing	Applying additional datasets	
DevOps	AWS Initialization	UI Prototype Security Audit	Gather UI Feedback	Host or Containerize
Evaluation	Research KM KPIs Research NLP Metrics	Apply NLP Experiments	Automated Testing	Human Preference (DPO)

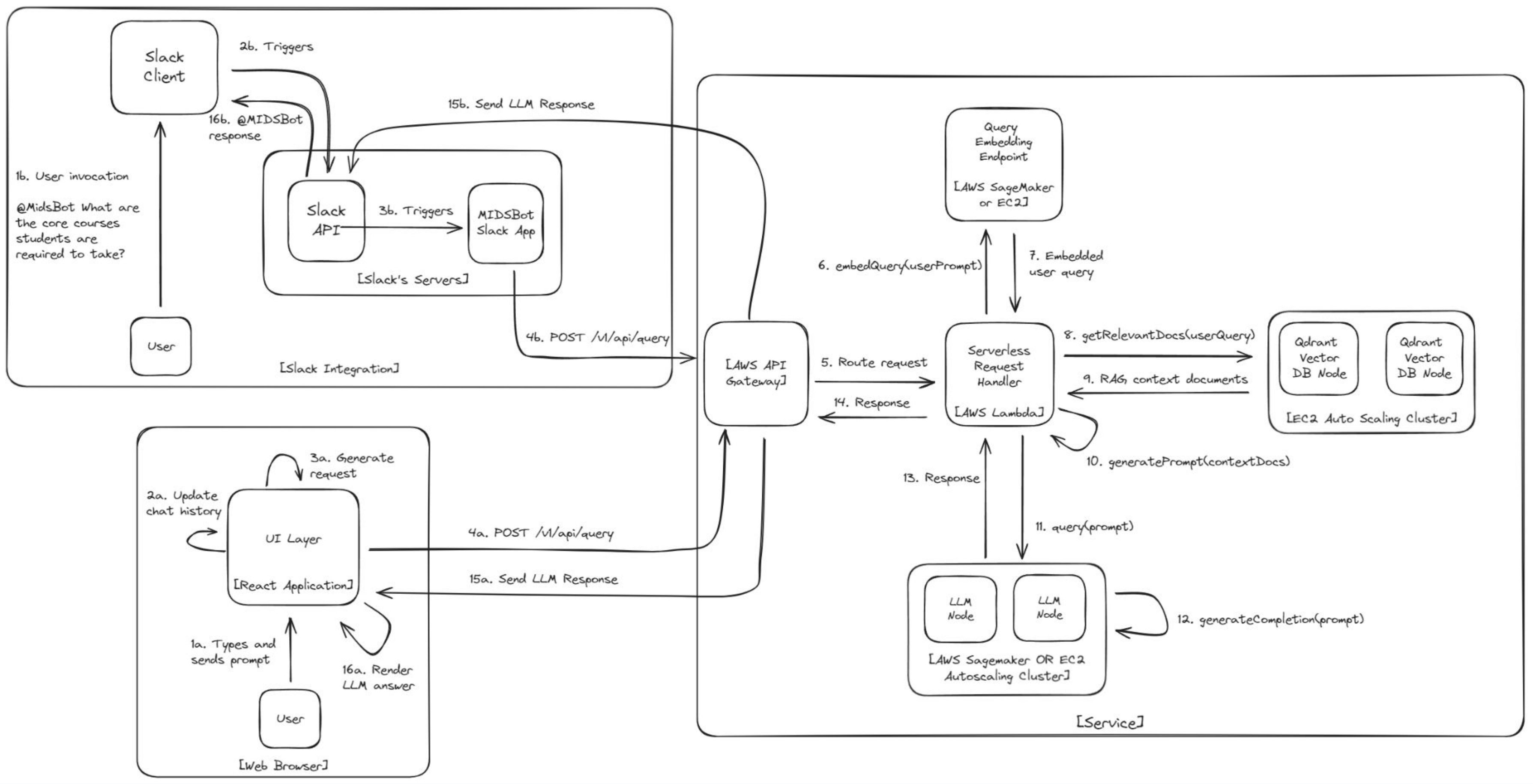
MVP



MIDS Chat

Enter your name

MVP Transition to Real Deal



Datasets

Data Collection

- **MIDS Intranet:** BeautifulSoup text scraping
- **Bcourses:** PDF processing MIDS Syllabi
- **Other publicly available sites** as needed
- **Slack**: admin export of #mids-class-rec channel
- L**LM Generated Question Answer Pairs**
  - Google Forms surveying students
  - Using openAi to generate questions AND answers based on available documents

Data Preparation

Data Cleansing

- Remove duplicate content (e.g., repetitive links to Facebook)
- Purge unnecessary content to maintain data relevance and reduce noise

Purge or anonymize all personal information attributes

Data Partitioning by

- Secure pages
- Public pages
- Course-specific pages

User Requirements:

Gather user input to understand their requirements, confirm our assumptions about features, user interface elements,

system usage and system functionality

- Google form
- Open Alpha and Beta to small group of students
- Human feedback via Direct Preference Optimization (DPO)

EDA

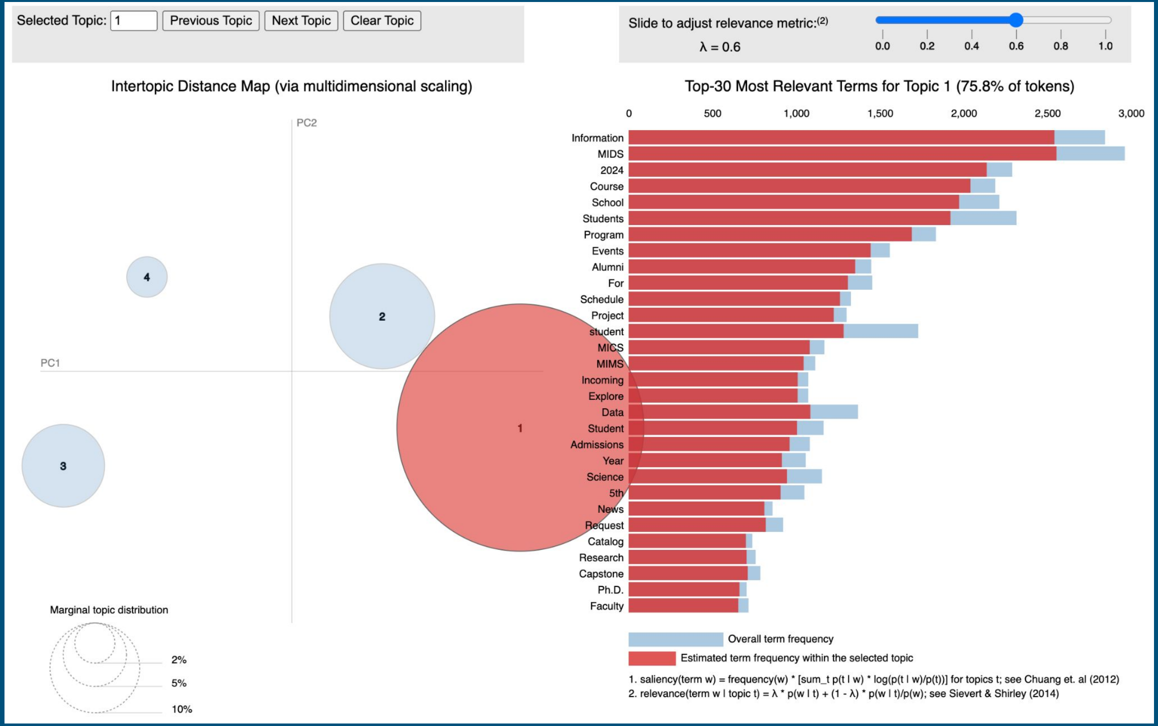
N-gram

Named Entity Recognition

Topic Modeling (LDA)

Topic Complexity

Sentiment Analysis



Selected Topic: 1

Slide to adjust relevance metric:(2)

λ = 0.6

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 1 (75.8% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Ethics and Privacy

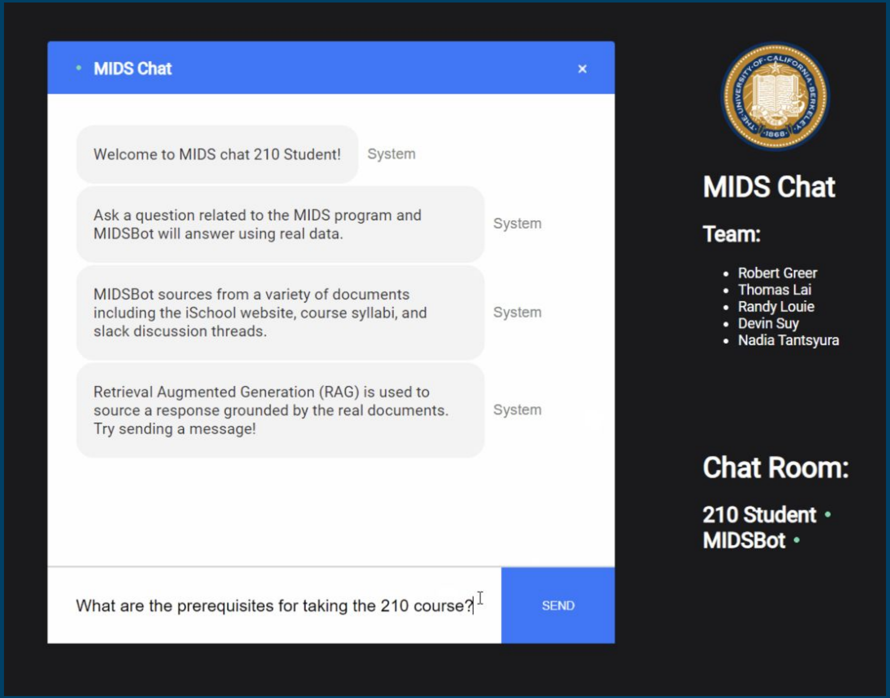
- Use of Slack
  - Terms and Conditions
- LLM considerations
  - Fine tuning on unbalanced dataset
  - Adversarial attacks
  - Guardrails
- Confidential data
  - PII
  - User sensitive information
  - User Privilege Classification on documents for access

Open Questions and Challenges

- RAG related
    - Generating Reference Answers
    - Model evaluation methods
    - Accessing Slack Data
- Deployment
    - MVP will be hosted locally with a static database
    - Can we build a generative system that run locally for an average users?
- A lot of cool ideas, what are the core features that make the cut?
    - Where will we draw the line ?
    - Not everyone can make varsity

Conclusion

Knowledge is found in many unstructured sources, we unlock it through a RAG-based approach.



Welcome to MIDS chat 210 Student! System

Ask a question related to the MIDS program and MIDSBot will answer using real data. System

MIDSBot sources from a variety of documents including the iSchool website, course syllabi, and slack discussion threads. System

Retrieval Augmented Generation (RAG) is used to source a response grounded by the real documents. Try sending a message! System

What are the prerequisites for taking the 210 course?

MIDS Chat

Team:

- Robert Greer
- Thomas Lai
- Randy Louie
- Devin Suy
- Nadia Tantsyura

Chat Room:

210 Student •
MIDSBot •

Notes

Pitching - do we keep SLM? Does that align with BErkeley

- Yes, it minimizes infrastructure and implementation complexity for Berkeley.
-