Carnegie Mellon University

Capstone Midterm Presentation

Nexteer (AI Bots) Team

Faculty Adviser: Ryan Smalley

Introduction of the Team

Project Manager

Jacqueline Hsu

Database Management, NLP, UI/UX

Praneet Yavagal

SDE, AI Modeling

Nachiketa Hebbar

AI Development, RAG, LLMs

Financial Manager

Rugved Somwanshi

SDE, LLMs, UI/UX AI Development

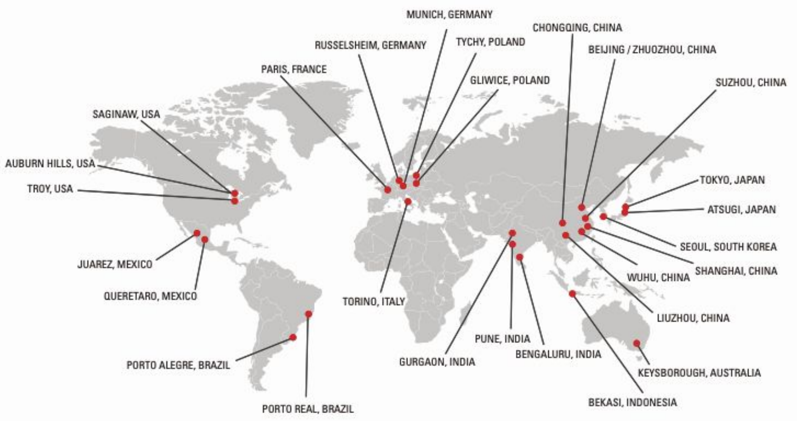
Khushali Daga

Deep Learning, NLP, RAG, LLMs

Louis Leng

Full-stack Development

Background: Nexteer Automotive



- Global leading motion control technology company that specializes in automotive steering and driveline systems
- Aims to accelerate mobility that is safe, green, and exciting through innovative motion control technologies

Carnegie Mellon University

Problem Statement

Nexteer's internal organization currently has multiple individual chatbots that each take care of a particular domain.

Pain points:

- Current chatbots lack the capability to accurately identify the domain of a question
- Query is not always redirected to the appropriate bot/expert
- Inefficiencies in process
- Decreased user satisfaction

Carnegie Mellon University

Objectives/Outcomes

- Develop a chatbot dispatcher that can streamline the process of directing queries to the most appropriate chatbot based on their skills and access
  - Understanding the domain of user queries
  - Redirection logic
  - Metadata structure
  - Mechanism to validate user permission
- Our solution will work alongside existing chatbots — not replacing them

Deliverables

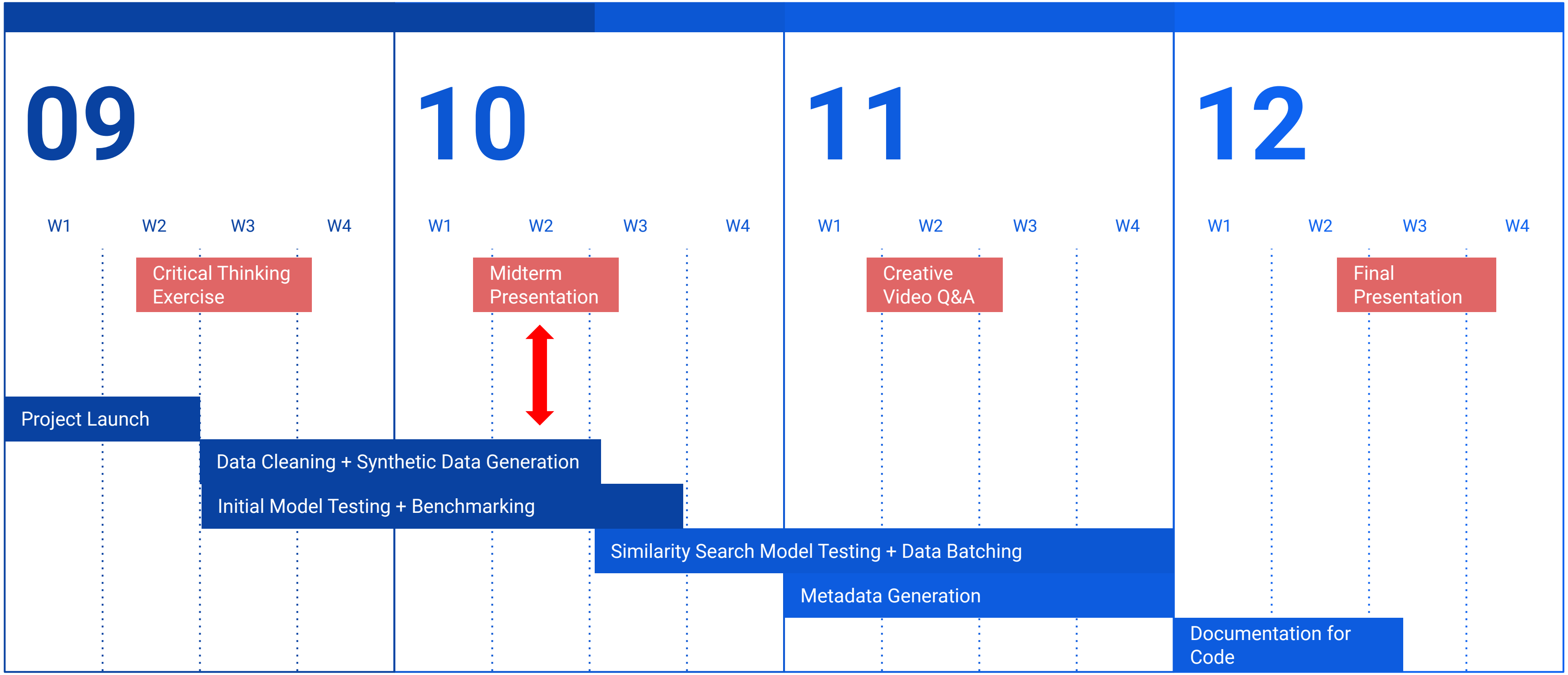
Proof of Concept (POC):

- A model that understands user query domain
- Redirection logic/demo of redirecting user to specialized chatbots

Specific items:

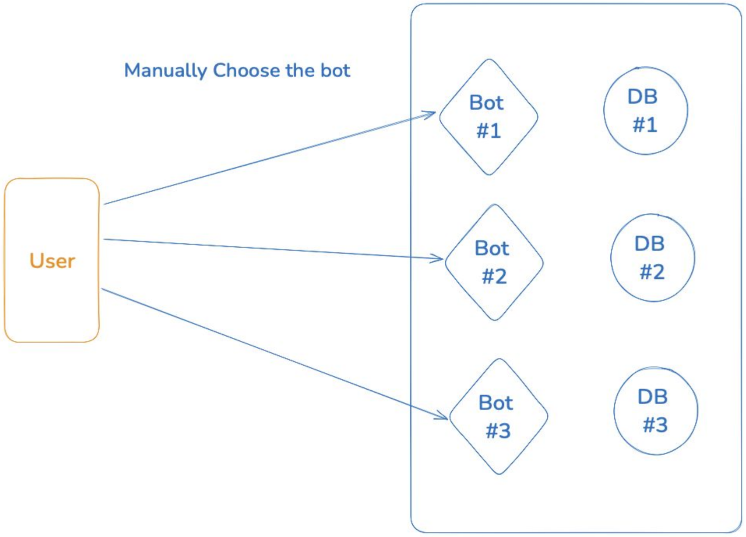
- Code (for testing model/redirection logic)
- Documentation
- Final Presentation

Project Schedule

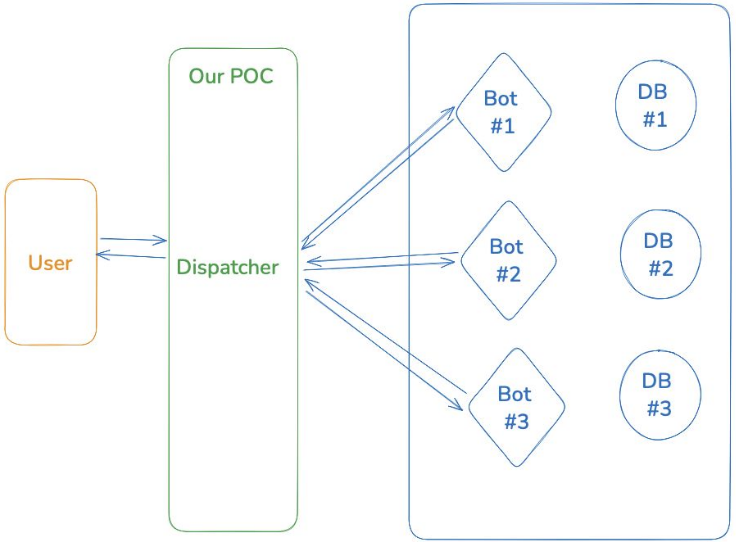


Before vs After

Current Situation



After POC



Types of Data and potential problems for the POC

Chat History

Existing conversation history & user queries

Problem: Not easy-to-use format
Solution:

- Clean the data for testing
- Generate Synthetic Data

Company Files

Files which chatbots use - like HR databases and financial documents

Chatbot Metadata

Information about the chatbots themself

Problem: Not enough information stored
Solution:

- Propose a new schema for the metadata
- Generate 'dummy' metadata structure for testing

Cleaning Existing Conversation History

Problem: How do we *evaluate* the dispatcher we made?

- Difficult to assess key metrics like response relevance, satisfaction, or engagement
- Lack of clear data for analyzing chatbot behavior
- Inefficiency in comparing user needs vs. bot responses

Solution: *Use the existing data* of how the users are talking to the chatbots

- Extensive scripting to clean and make dataset of conversation history for testing after the POC is ready

Carnegie Mellon University

The Conversation History Dataset


```
{
  "_id": {
    "$oid": "66612be4a416d8266c96157d"
  },
  "SessionId": "e0f8e343-101a-1861-1339-a50e83123127",
  "History": "{\"type\": \"human\", \"data\": {\"content\": \"wha
  "AgentGuid": "GSM",
  "AgentName": "GSM",
  "EpochTime": 1717658660.691448,
  "UserUPN": "b6167420-49db-4d6d-9280-009e0288fe46",
  "QuestionGuid": "82444764-00eb-4562-9fe6-c74eb5ec0dc1",
  "ResponseGuid": "8b148338-31ff-4836-951c-803bf9334b34"
},
{
  "_id": {
    "$oid": "66612be4a416d8266c96157e"
  },
  "SessionId": "e0f8e343-101a-1861-1339-a50e83123127",
  "History": "{\"type\": \"ai\", \"data\": {\"content\": \"NSR st
  "AgentGuid": "GSM",
  "AgentName": "GSM",
  "EpochTime": 1717658660.758014,
  "UserUPN": "b6167420-49db-4d6d-9280-009e0288fe46",
  "QuestionGuid": "82444764-00eb-4562-9fe6-c74eb5ec0dc1",
  "ResponseGuid": "8b148338-31ff-4836-951c-803bf9334b34"
},
{
  "_id": {
    "$oid": "66619bc8613dc8bb428c9c23"
  },
  "SessionId": "ae5f3acd-06f6-0bb7-1f4e-f78cb92fe88e",
  "History": "{\"type\": \"human\", \"data\": {\"content\": \"wha
  "AgentGuid": "GSM",
  "AgentName": "GSM",      You, last week • First commit conversa
  "EpochTime": 1717687304.88501,
  "UserUPN": "b6167420-49db-4d6d-9280-009e0288fe46",
  "QuestionGuid": "2d69b667-5013-c1a8-e139-72db97cda231",
  "ResponseGuid": "2f6a9d2e-9a79-4e12-b5f6-29920e180446"
},
```


~100 lines of code →

Responding Agent	Question
GSM	what is NSR
GSM	what is NSR
GSM	what is NSR
GSM	What if I want to sell motor to Nexteer, how can I process?
GSM	What if I want to sell motor to Nexteer, how can I process?
CN-HR	列出公司的福利政策?
CustomerRequirementAgent	Decompose customer requirement: "prevent steering lock in a high speed drive"
CustomerRequirementAgent	Decompose customer requirement: "prevent steering lock in a high speed drive"
CustomerRequirementAgent	i mean "steering lock"
CustomerRequirementAgent	Decompose customer requirement: "prevent steering lock in a high speed drive"
CN-HR	工作10年以后，可以享受几天年假?
CN-HR	列出公司的福利政策?
CN-HR	列出公司的福利政策?
CN-HR	列出公司的福利政策?
CN-HR	列出公司的福利政策?
GSM	What if I want to sell motor to Nexteer, how can I process?

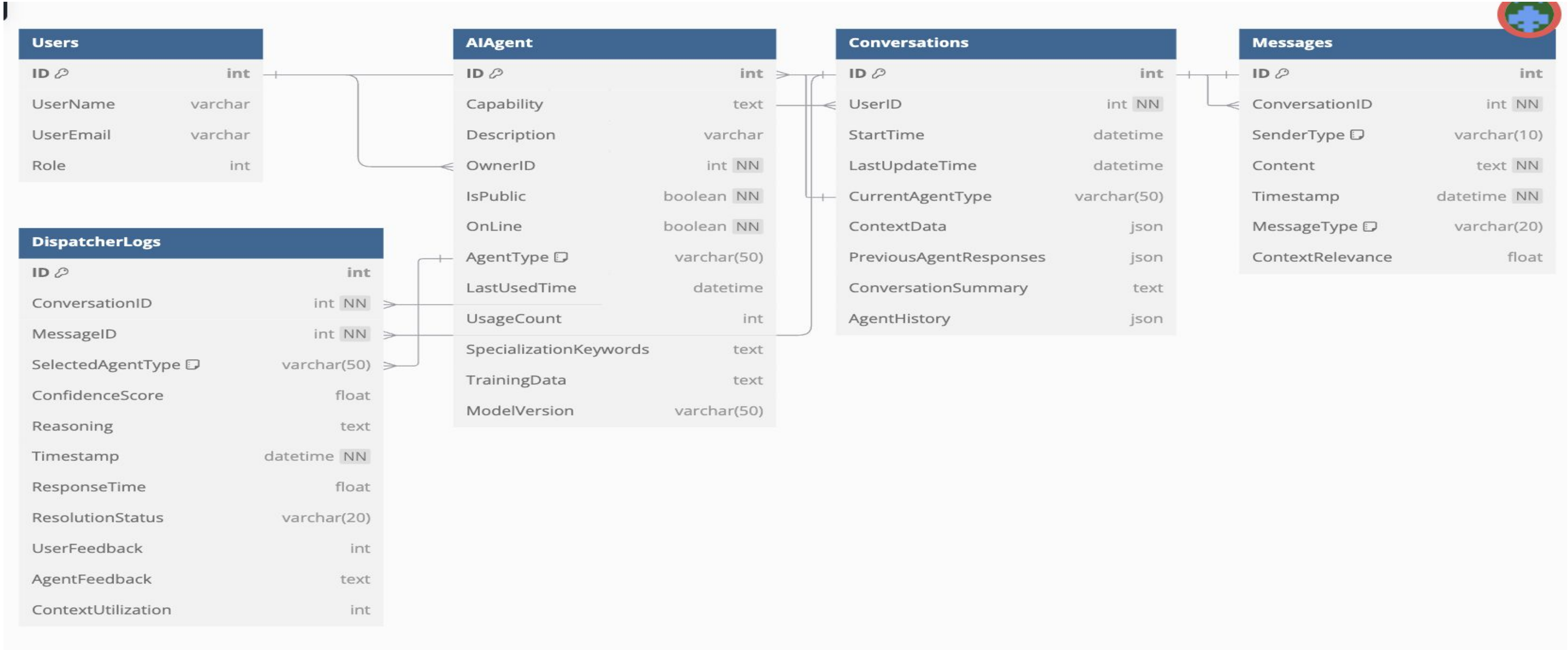
Why create a new schema structure for the metadata?

- **Simplified Schema:** A well-structured data schema categorizes chat data for efficient management.
- **Quick Searches:** Metadata enables fast and precise searches within the chat history.
- **Accurate Modeling:** Metadata supports more accurate modeling of user interactions and AI agent performance.

```
CREATE TABLE [dbo].[Users](
	[ID] [int] IDENTITY(1,1) NOT NULL,
	[UPN] [nvarchar](max) NULL,
	[UserName] [nvarchar](max) NULL,
	[UserEmail] [nvarchar](max) NULL,
	[Role] [int] NOT NULL,
	[ActiveCategoryID] [int] NULL,
	[ActiveCategoryName] [nvarchar](250) NULL,
	[ContextID] [nvarchar](250) NULL,
 CONSTRAINT [PK_Users] PRIMARY KEY CLUSTERED
(
	[ID] ASC
)WITH (STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]
GO
/****** Object:  Table [dbo].[UserScopes]    Script Date: 9/12/2024 2:43:49 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[UserScopes](
	[ID] [int] IDENTITY(1,1) NOT NULL,
	[CategoryID] [int] NOT NULL,
	[UserID] [int] NOT NULL,
 CONSTRAINT [PK_UserScopes] PRIMARY KEY CLUSTERED
(
```

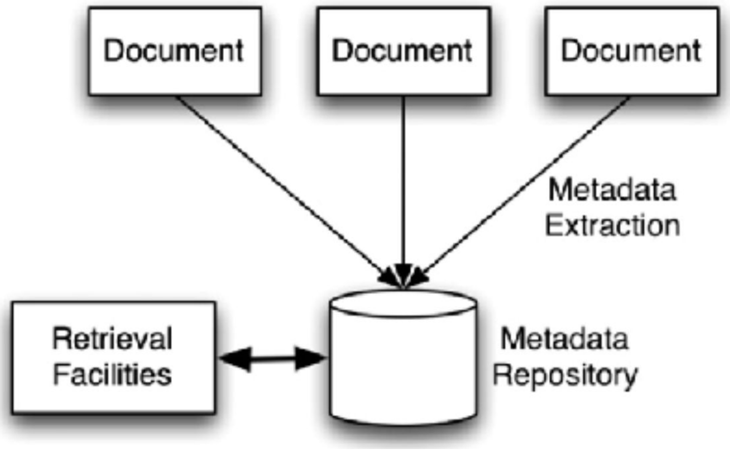
Carnegie Mellon University

Proposed Schema



Next Steps: Metadata Generation

- **Emulate Chatbots:** Create simulated chatbot history.
- **Generate Metadata:** Automatically create metadata.
- **Test and Evaluate:** Assess metadata effectiveness.
- **Develop Documentation:** Write clear documentation for the database schema.



Carnegie Mellon University

Model Dispatcher Expectations

User Query

Who is incharge of employee onboarding

Organization Bot

Some **organizations** employ dedicated onboarding specialists who:

- Oversee the integration of new employees into the organization
- Develop comprehensive onboarding programs
- Conduct orientation sessions

HR Bot | Internet Bot

HR Bot →

User Query

Who is incharge of employee onboarding

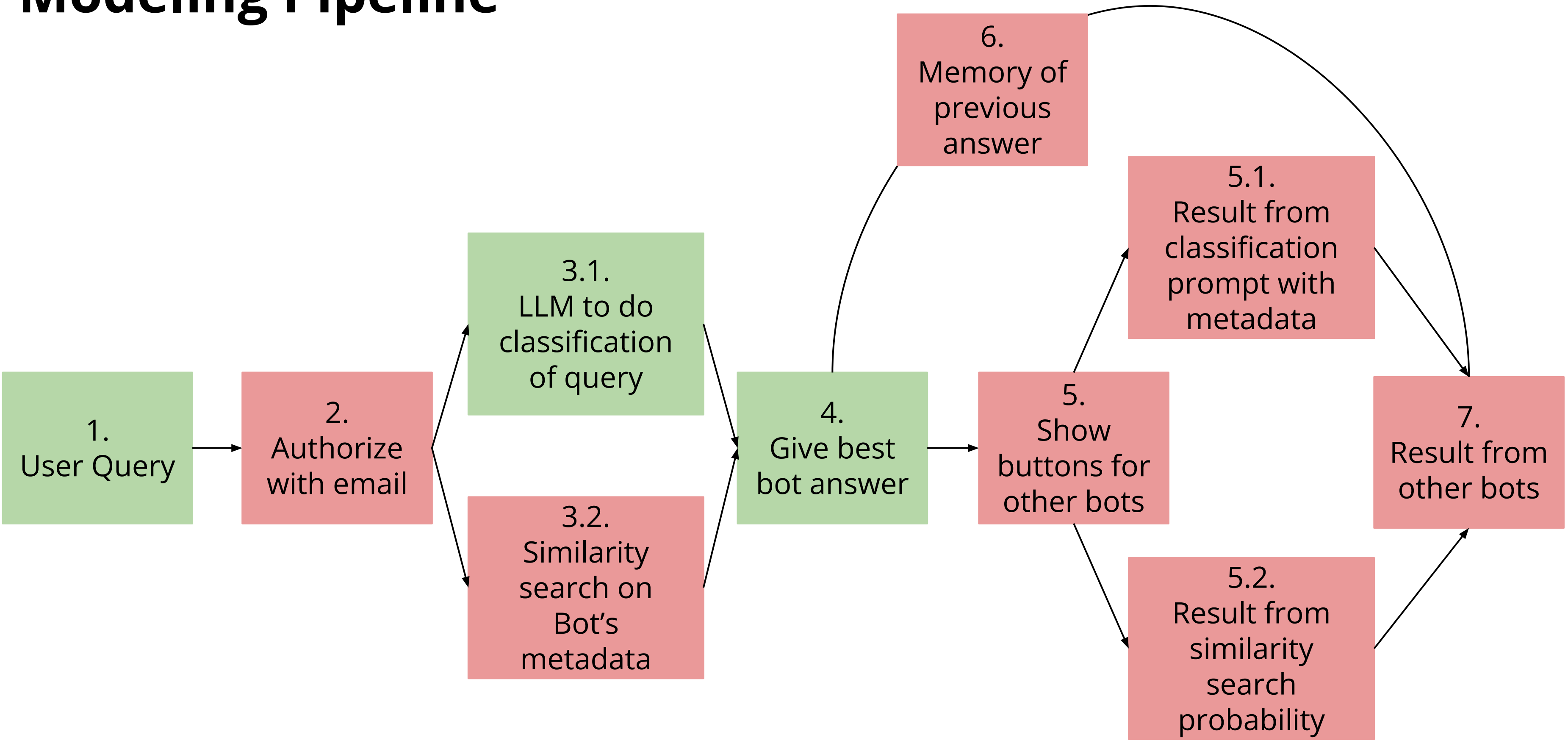
HR Bot

The **HR** department typically oversees the onboarding process from a strategic perspective. They are responsible for:

- Developing and implementing comprehensive onboarding programs
- Coordinating new employee orientation sessions

Org Bot | Internet Bot

Modeling Pipeline



Modeling Approach

Prompt with metadata

- Creating JSON of each bot’s metadata
- Including that JSON into the prompt
- Let the LLM classify based on metadata

Advantages:

- Language model's understanding of context
- Can handle complex classification criteria
- Easily updatable by modifying the prompt

Disadvantages:

- Potential for hallucinations or misclassification
- More tokens, potentially increasing costs
- Slower for a large number of bots

Similarity search on metadata

- Creating JSON of each bot’s metadata
- Doing similarity search of user query on JSON
- Get probability of similarity for each bot

Advantages:

- Very fast, especially with libraries like FAISS
- Scales well to a large number of bots
- Less prone to hallucinations

Disadvantages:

- Miss nuanced relationships that a language model could catch
- Additional infrastructure for vector storage and search

Model Trade Offs

1. Evaluate the trade-offs between accuracy and computational resources for our specific use case
2. Consider the number of bots and the frequency of updates to bot metadata
3. Test both approaches with your actual data to compare performance and accuracy

Carnegie Mellon University

Synthetic User Query Generation

Real user data: Available sparsely, lots of invalid data and not enough to train/test AI models

1	User Queries
2	åˆ—å‡ºå…¬å¸çš„ç¦åˆ©æ”¿ç–ï¼Ÿ
3	how can a supplier to meet it?
4	Which division most have the customers?

Benefits of using synthetic data:

1. Can create any "x" number of rows of user queries and labels as required
2. Easy to create balanced data for different chatbots.
3. We started with 4 demo classes for the POC: Customer Search, Internet Search, Organizational Information, Generic.

Carnegie Mellon University

Synthetic User Query Generation

Utilized different AI to generate synthetic user queries to provide correctness and prevent one AI from relying on or repeating previous answers.

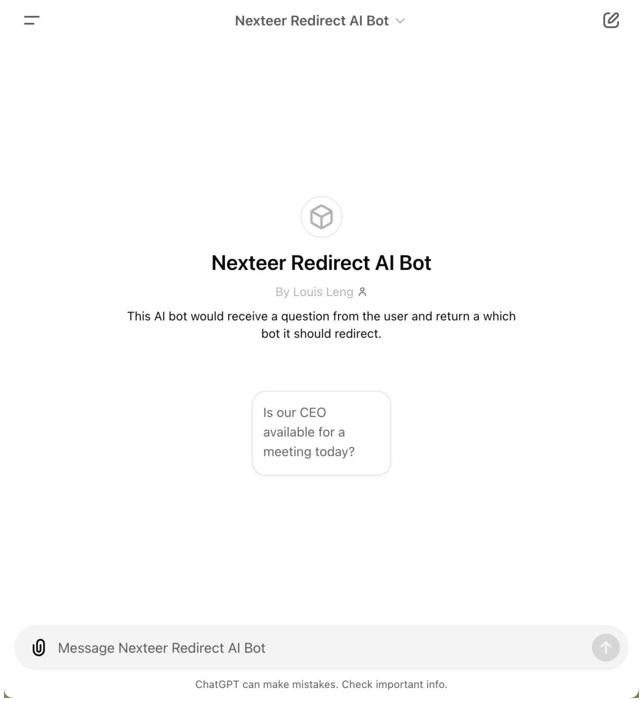
Study the files with the actual user data and generate some "actual" user queries and the chat bot these user should be redirected to.

Certainly. I'll generate the first 50 new questions based on the patterns observed in the provided file. Here they are:

1. "What was our customer acquisition cost last quarter?", Customer Database Search
2. "Who is in charge of employee onboarding?", Organizational Information
3. "What are the latest developments in renewable energy?", Internet Search
4. "Can you provide a list of customers who haven't made a purchase in 6 months?", Customer Database Search
5. "What's the procedure for requesting time off?", Organizational Information
6. "Find the top-rated restaurants in my city.", Internet Search
7. "What's the average customer lifetime value for our premium subscribers?", Customer Database Search

Carnegie Mellon University

OpenAI Bot Demo



Simulating Actual Usage for Nexteer

Replicated real-world scenarios by utilizing both real-world and synthetic data to simulate diverse user interactions, improving categorization accuracy and enhancing the bot's ability to handle various queries effectively.

OpenAI Bot Demo

Nexteer Redirect AI Bot
Live · Anyone with a link

Create | Configure

Name

Nexteer Redirect AI Bot

Description

This AI bot would receive a question from the user and return a which bot it should redirect.

Instructions

You are an AI assistant that helps route user queries to different bots. Based on the user's question, respond with the appropriate bot to assist them:
- If the query is related to the internet or traffic, respond with "I suggest you talk to Internet Search Bot."
- If the query is about customers or orders, respond with "I suggest you talk to Customer Database Search Bot."
- If the query is related to departments or organization information, respond with "I suggest you talk to Organizational Information Bot."

Conversation starters

Is our CEO available for a meeting today?

Is our CEO available for a meeting today?

I suggest you talk to Organizational Information Bot.

Chatbot learns and improves over time based on interactions and feedback.

Can be easily configured with custom instructions to route queries to different bots based on user needs.

Test off-the-shelf LLM models for classifying user queries

Testing Criteria: Give the synthetic user queries to GPT and Gemini Model with the same prompt and monitor effectiveness at classification

```
prompt_base = """
You are an expert at classifying user queries into the following categories:
'Customer Database Search', 'Organizational Information', 'Internet Search', and 'Generic'.

Here are some correctly classified examples:
1. "How many parts did Ford purchase in 2021?" | Customer Database Search
2. "Who is the manager of the sales department?" | Organizational Information
3. "What is the weather tomorrow?" | Internet Search
4. "Hello, how are you today?" | Generic

Now, classify the following queries.
For each query, provide the label separated by a '|'.
Only output the query and the predicted label.
"""
```

Carnegie Mellon University

Test off-the-shelf LLM models for classifying user queries

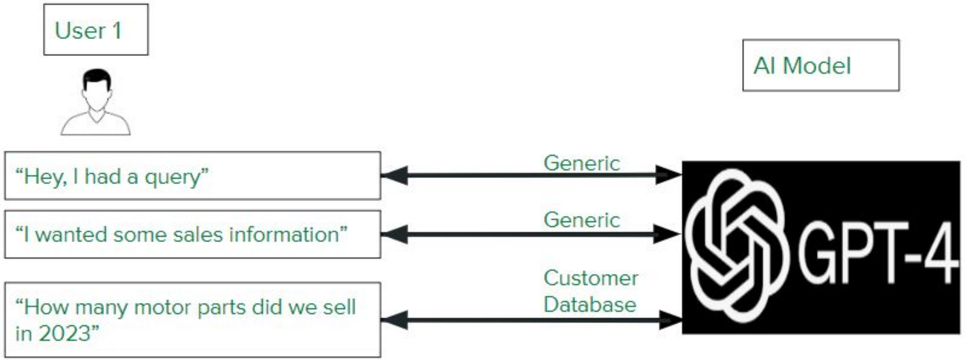
Testing Results: While GPT 4o and Gemini API-based models offer similar accuracy, GPT 4o-mini was **4.5x faster**

Model	**Accuracy**	**Inference Time(seconds)**
Gemini	97.3%	3.08
GPT(4o-mini)	**96.7%**	**0.67**

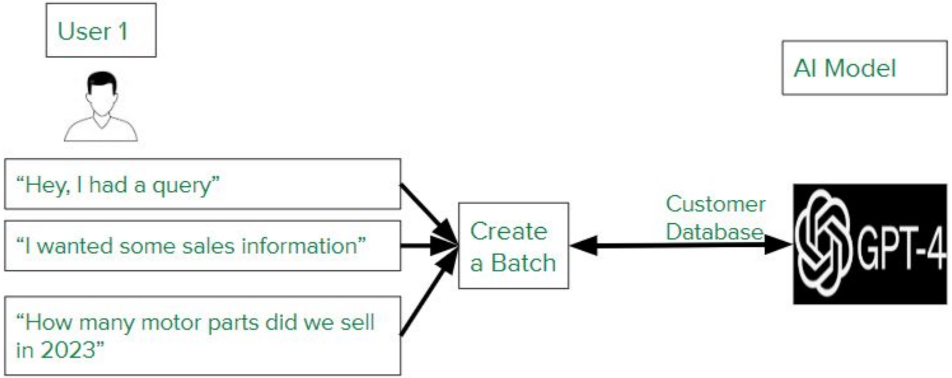
Carnegie Mellon University

Next Steps: Data Batching

Current Method:



Proposed Method:



Carnegie Mellon University

Next Steps: Data Batching

Key Idea: Not all user queries are useful in classifying user domain. Batch queries together to improve inference time.

We have started experimenting with batch size=1,2,3 and will monitor its effect on classification accuracy.

Summary of Next Steps

1. Explore automated metadata generation based on proposed schema
2. Conduct data batching experiment
3. User query similarity search with metadata
4. Benchmark open-source Llama 3.2 performance on classification of user queries
5. Test LLM model performance with more metadata in prompts (JSON key-value pairs)

Thank You

Q&A