

MICHIGAN STATE

UNIVERSITY

Beta Presentation

AI System Testing Framework

The Capstone Experience

Team Ally

Vu Ho

Andrew Dagher

Gabe Moraru

Michael Plante

Ethan Gomez

Amit Wagh

Department of Computer Science and Engineering
Michigan State University

Spring 2025



*From Students...
...to Professionals*

Project Overview

- GenAI Models Behave Unpredictably
- Testing Framework To Evaluate LLM Performance
- Uses a “Solution-First” Comparison
- Helps Identify Good/Poor Use Cases



Team Member's Technical Tasks

Technical Tasks Assigned

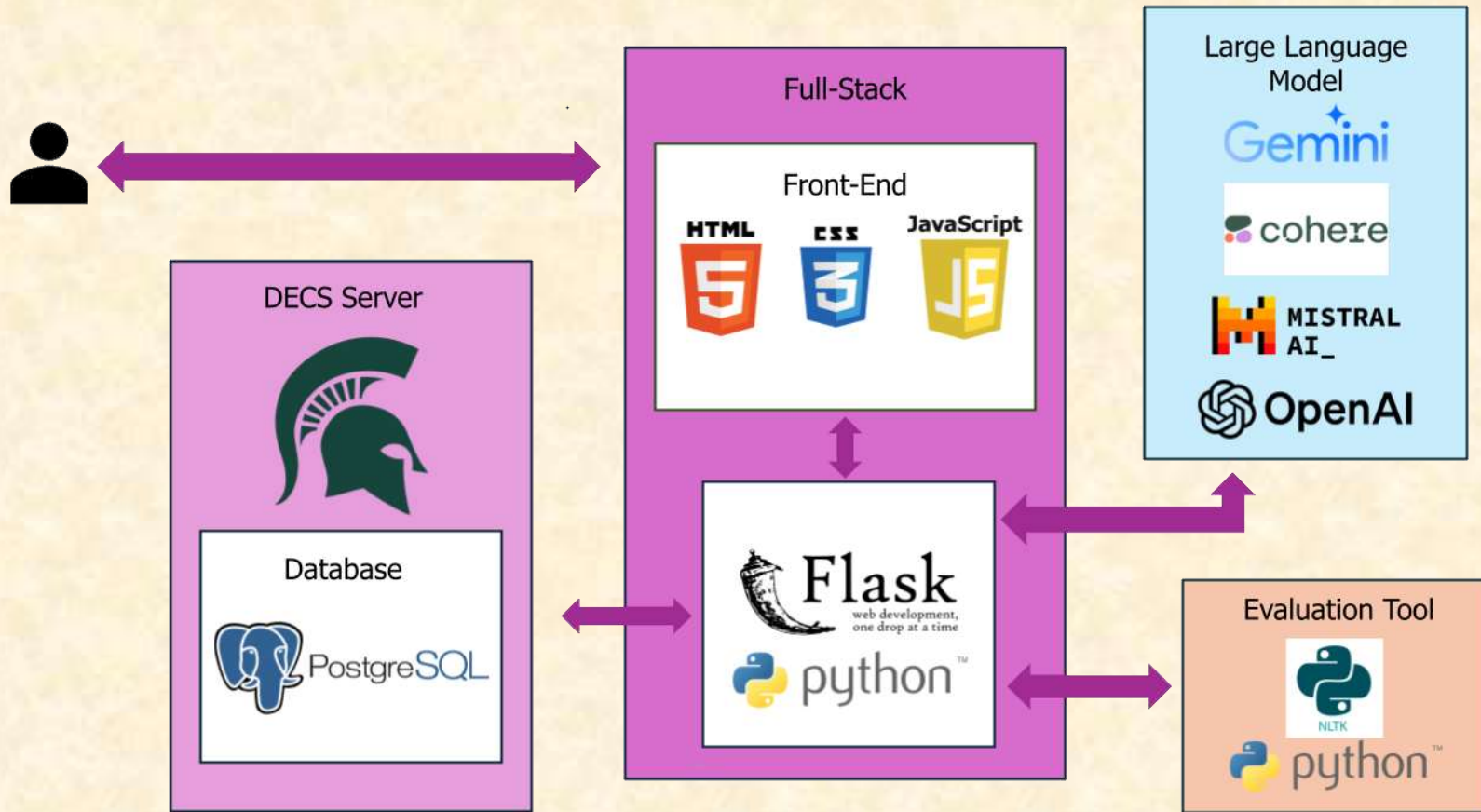
- Andrew Dagher
 - Added ability to select different models for evaluation
 - Develop configure metrics feature
 - Develop AI generated reference feature
- Michael Plante
 - Multiple conversations and multiple references
 - Uploads and downloads throughout application
 - Benchmarking and conversation analysis pages
- Ethan Gomez
 - Developed Coherence, Fluency, and Conciseness metrics
 - Developed Redundancy and merged into Coherence
 - Created metric API and Documentation
- Amit Wagh
 - Improve accuracy metric
 - Create relevancy and bertscore metrics
 - Implement AI analysis
- Gabe Moraru
 - Create frontend for most pages (chat, evaluation, etc)
 - Develop integration UI and backend
 - Implement admin/login features using database
- Vu Ho
 - Implemented completeness and Rouge-N metrics
 - Implemented AI reference dropdown menu
 - Added OpenAI and Mistral models

Technical Tasks Completed

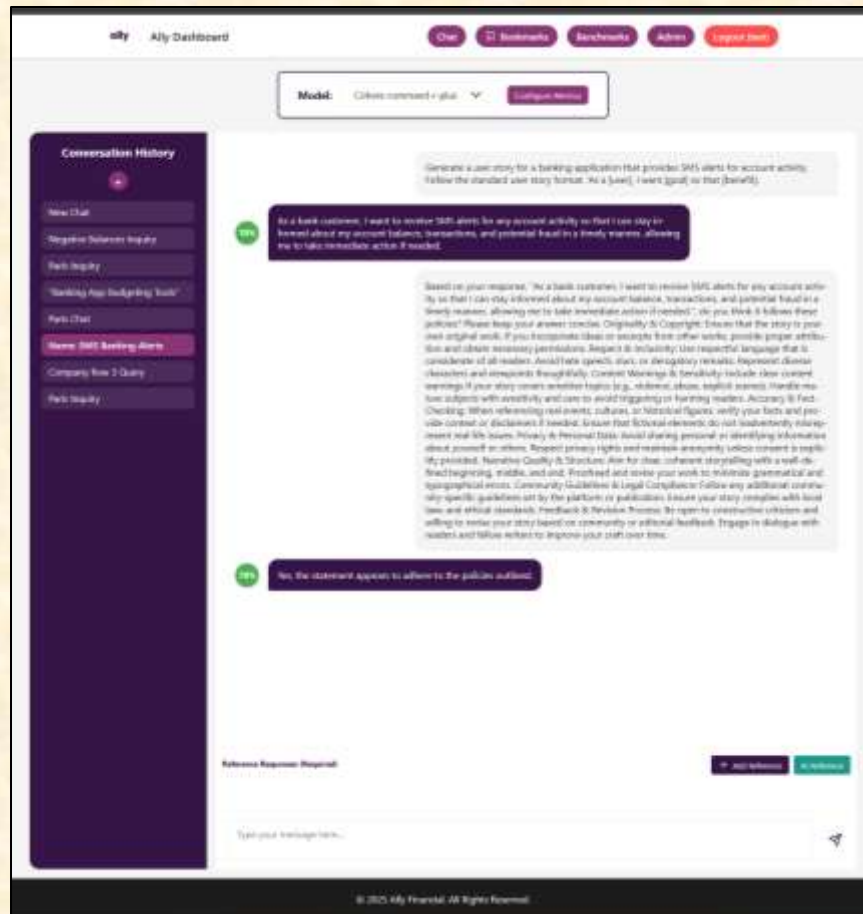
- Andrew Dagher
 - Added ability to select different models for evaluation
 - Develop configure metrics feature
 - Develop AI generated reference feature
- Michael Plante
 - Multiple conversations and multiple references
 - Uploads and downloads throughout application
 - Benchmarking and conversation analysis pages
- Ethan Gomez
 - Developed Coherence, Fluency, and Conciseness metrics
 - Developed Redundancy and merged into Coherence
 - Created metric API and Documentation
- Amit Wagh
 - Improved accuracy metric
 - Created relevancy and bertscore metrics
 - Implemented AI analysis
- Gabe Moraru
 - Created frontend/JavaScript for most pages
 - Developed integration for UI and backend
 - Implemented admin/login features using database
- Vu Ho
 - Implemented completeness and Rouge-N metrics
 - Implemented AI reference dropdown menu
 - Added OpenAI and mistral models



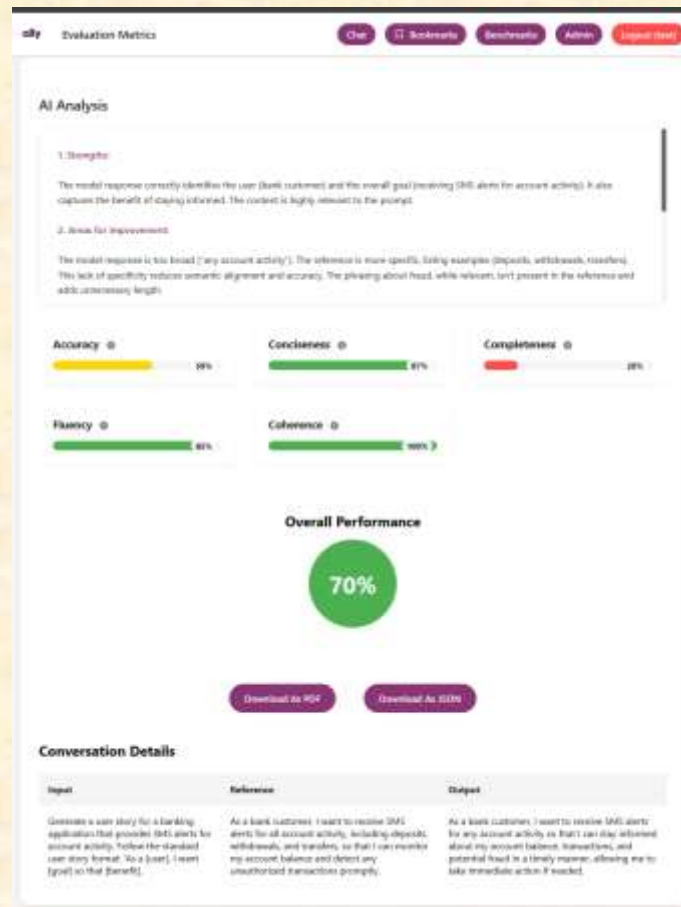
System Architecture



Chat Dashboard



Evaluation Dashboard



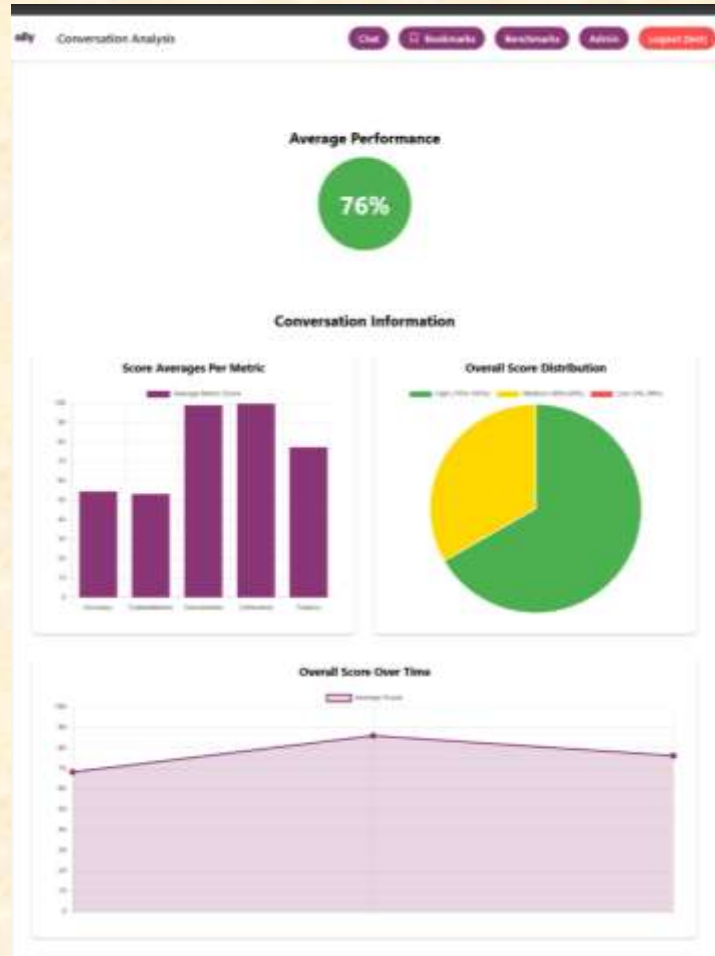
Bookmarks Page

The screenshot shows the 'ally Bookmarks' page. At the top, there is a navigation bar with the 'ally' logo and the title 'Bookmarks'. To the right of the title are five buttons: 'Chat', 'Bookmarks' (highlighted), 'Benchmarks', 'Admin', and 'Logout (test)'. Below the navigation bar is a light blue informational banner that reads: 'Bookmarks are shared among all users. This page displays all bookmarked evaluations across the system.' Below the banner is a table with the following columns: 'Score', 'Prompt', 'Reference', 'Response', 'Model', and 'Bookmarked By'. The table contains three rows of data. The first row has a score of 68% (yellow circle), a prompt 'Can you tell me why this ans...', a reference 'The provided answer adhere...', a response 'This answer adheres to the p...', a model 'Cohere command-r-plus', and 'test' bookmarked by. The second row has a score of 72% (green circle), a prompt 'Can you tell me what people ...', a reference 'The people that have negativ...', a response 'The following people have n...', a model 'Google Gemini 1.5 Flash', and 'test' bookmarked by. The third row has a score of 70% (green circle), a prompt 'Generate a user story for a b...', a reference 'As a bank customer, I want t...', a response 'As a bank customer, I want t...', a model 'Cohere command-r-plus', and 'test' bookmarked by. At the bottom of the table is a horizontal scrollbar.

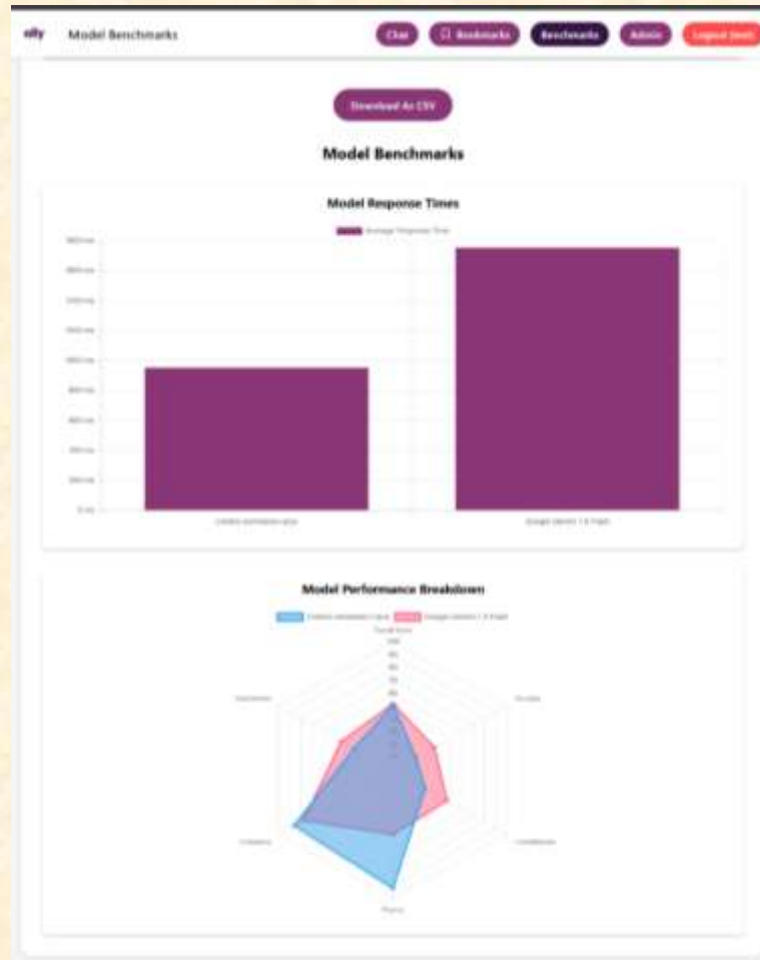
Score	Prompt	Reference	Response	Model	Bookmarked By
68%	Can you tell me why this ans...	The provided answer adhere...	This answer adheres to the p...	Cohere command-r-plus	test
72%	Can you tell me what people ...	The people that have negativ...	The following people have n...	Google Gemini 1.5 Flash	test
70%	Generate a user story for a b...	As a bank customer, I want t...	As a bank customer, I want t...	Cohere command-r-plus	test



Conversation Analysis Page



Benchmarks Page



What's left to do?

- Features
- Stretch Goals
 - Expand API for Project
- Other Tasks
 - Stylistic Changes
 - Optimization
 - Bug Fixes
 - Improve Documentation & Add ReadMe

Questions?

?

?

?

?

?

?

?

?

?

