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

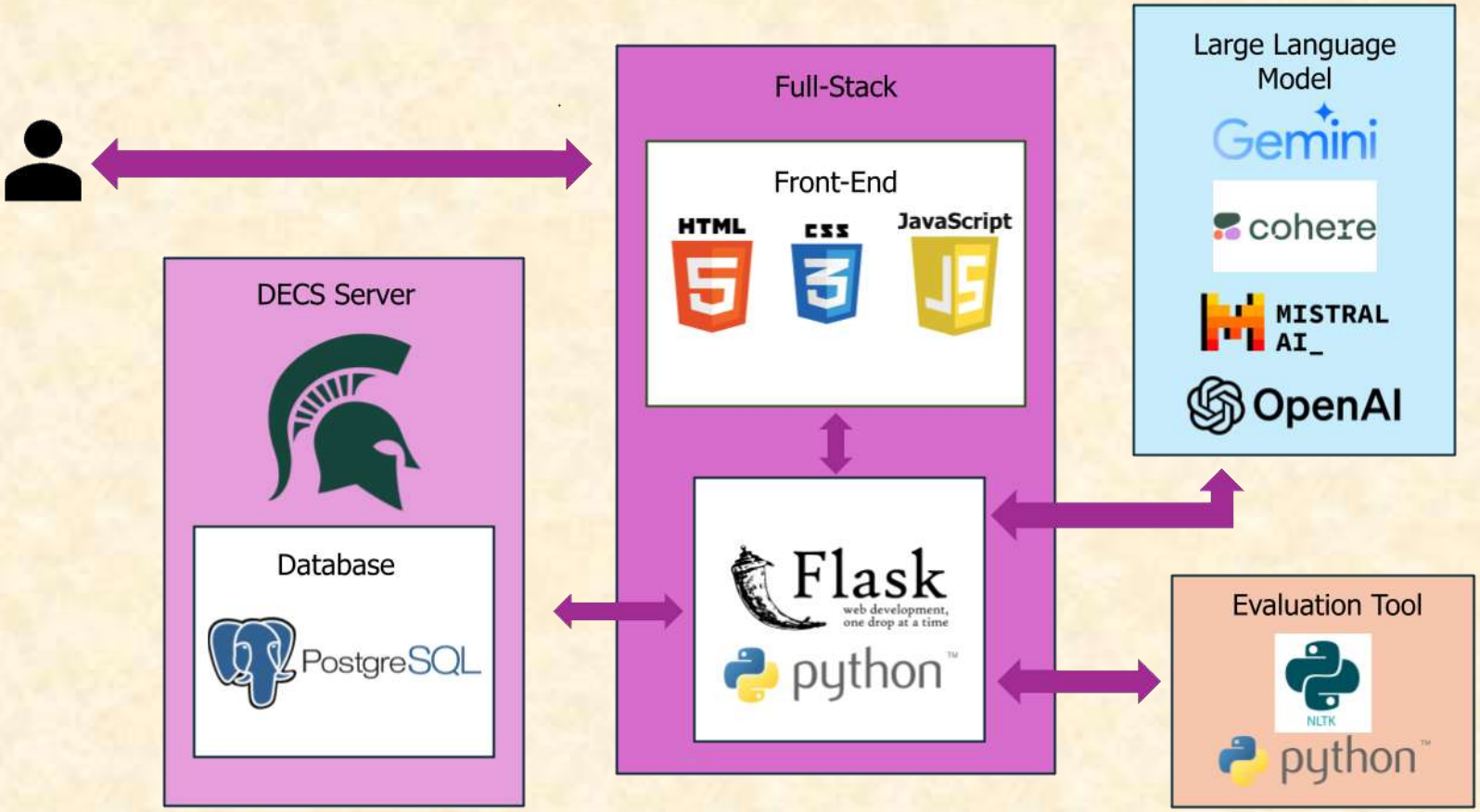
Technical Tasks Assigned

- 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

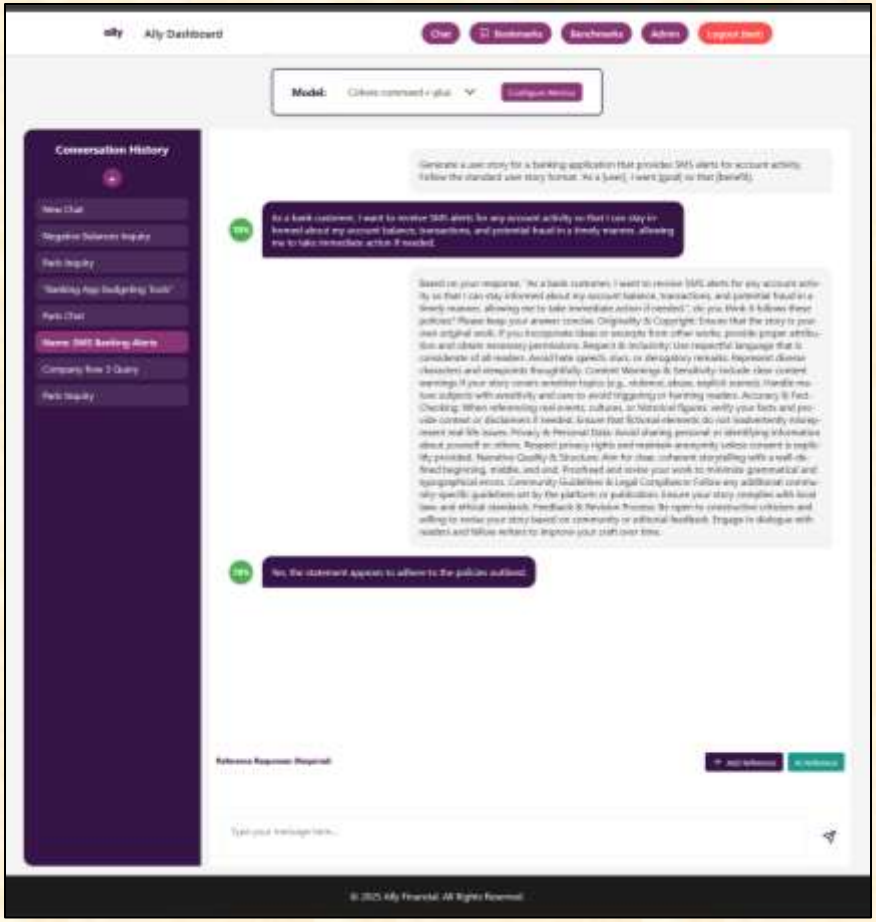
Technical Tasks Completed

- 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

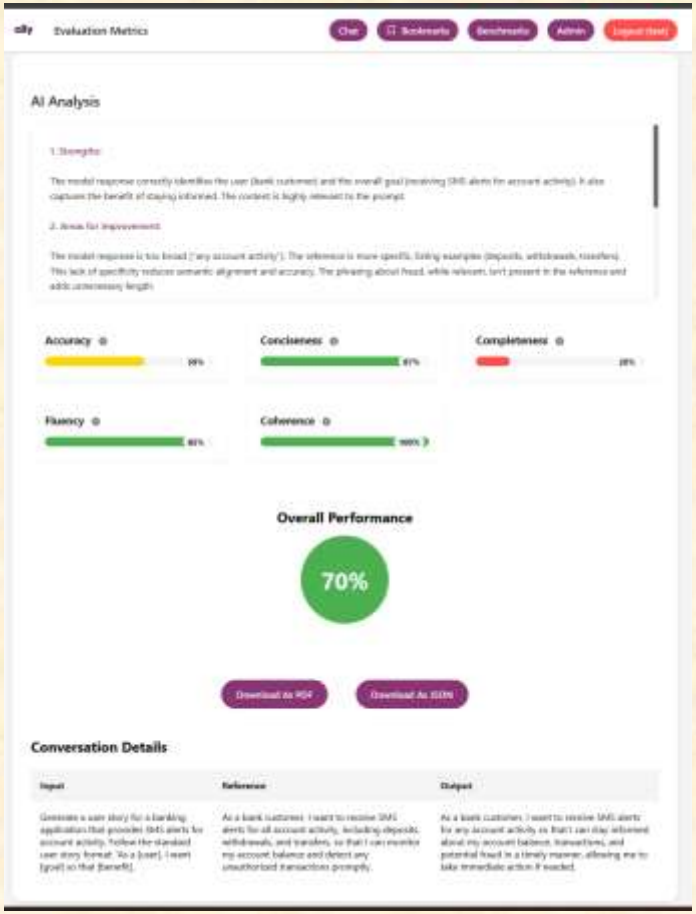
System Architecture



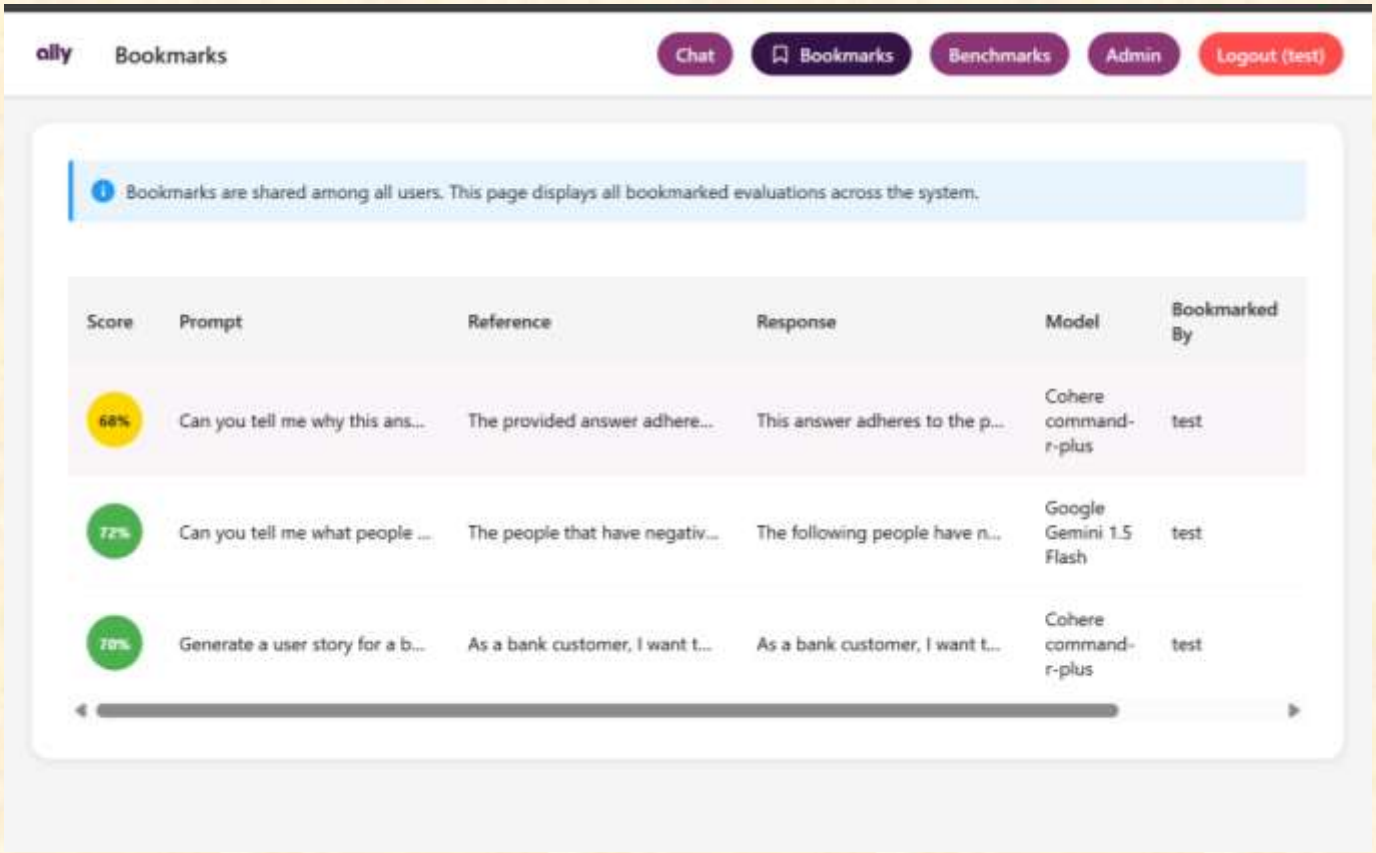
Chat Dashboard



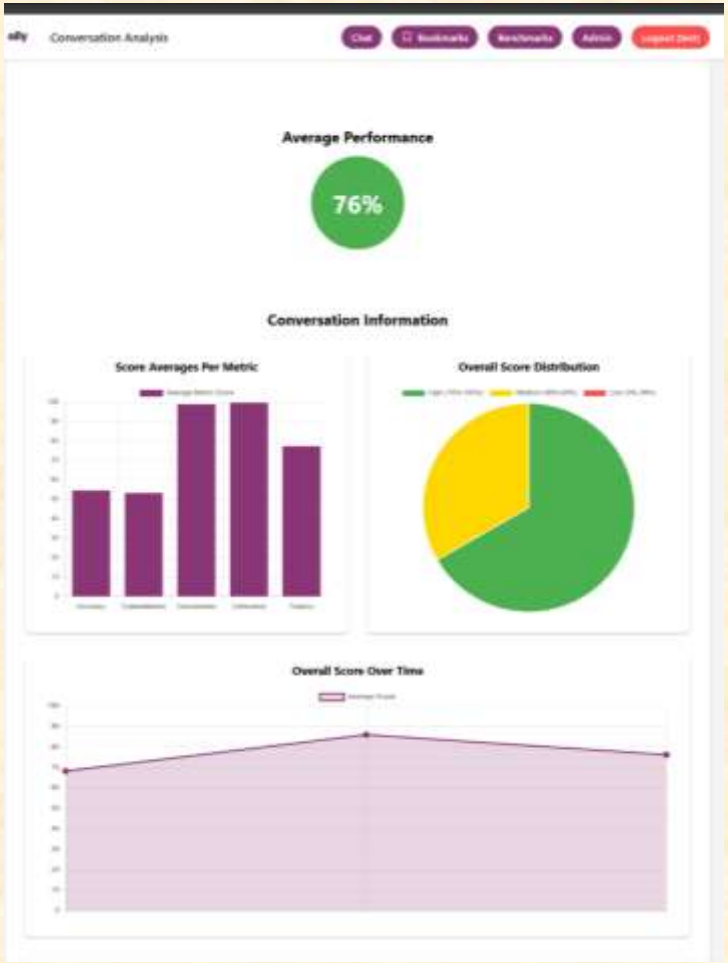
Evaluation Dashboard



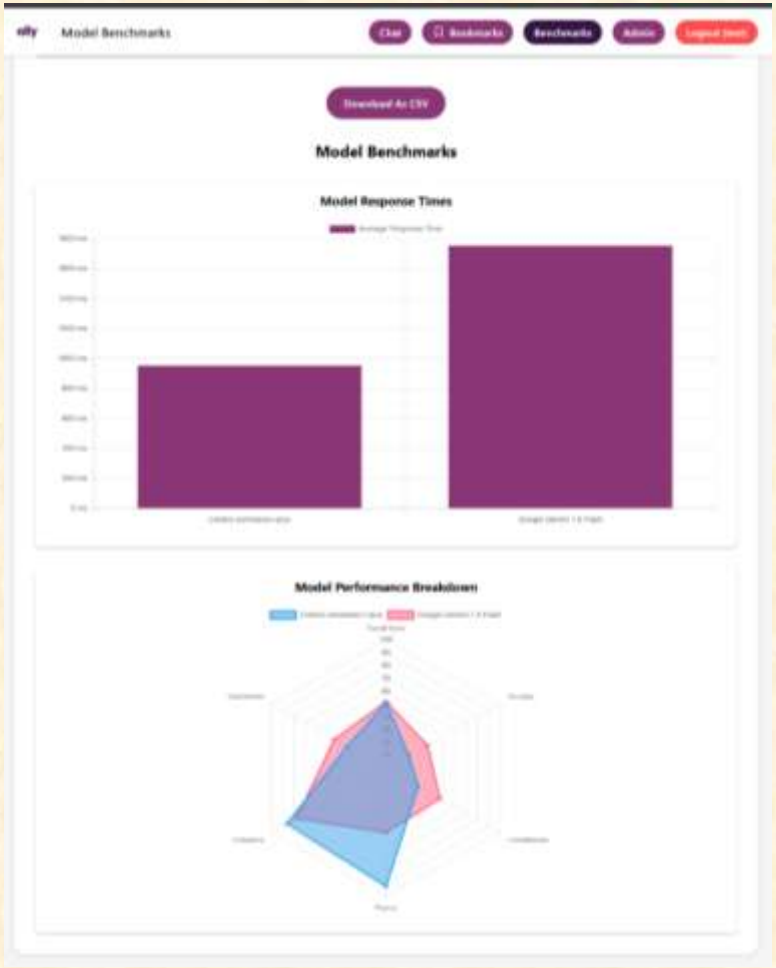
Bookmarks Page



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?