AI Medical Benchmark

Week 1 Individual Project Report

Evaluating AI Performance on Medical Licensing Exams

Project Overview

Problem: Which AI model performs best on medical exams?

Why Important:

- AI usage in healthcare is rapidly growing
- Medical students need reliable study assistants
- Critical to identify which models have the best medical knowledge depth

Solution: Create a comprehensive AI benchmark system

Core Features

🎯 Four Main Components

1. **Question Dataset** - Standardized USMLE questions
2. **Model Integration** - Multi-provider AI framework
3. **Evaluation System** - Automated answer checking
4. **Results Dashboard** - Performance visualization

Project Scope

📊 By the Numbers

Features

- **4** Core Features
- **4** Stretch Features
- **8** Total Features

Requirements

- **5** Core User Stories
- **3** Stretch User Stories
- **8** Total Requirements

Feature 1: Question Dataset

Problem

USMLE questions aren't available in structured, machine-readable format

Solution

- Scrape official Step 1 Free 120 questions
- Store questions, answer choices, and correct answers
- Capture associated medical images
- Ensure uniform JSON/CSV format

Feature 2: Model Integration

Problem

Different AI models use different APIs and formats

Solution

- Unified framework for multiple AI providers
- Support OpenAI, Anthropic, Google, etc.
- Standardized interface using API keys and model IDs
- Graceful handling of rate limits and errors

Feature 3: Evaluation System

Problem

Need automated way to verify AI answers

Solution

- Parse AI responses to extract chosen answers
- Compare against correct answers automatically
- Track correct/incorrect responses
- Store raw responses for detailed analysis

Feature 4: Results Dashboard

Problem

Raw model outputs are hard to interpret at scale

Solution

- Display accuracy percentage per model
- Show per-question logs with responses
- Filter results (correct/incorrect/by model)
- Provide charts and graphs for comparison

Stretch Features

🚀 Advanced Capabilities

- **Image Handling** - Test with/without medical images
- **Prompt Perturbation** - Test consistency with question variations
- **Lab Value Expansion** - Add clinical data for deeper reasoning
- **Advanced Metrics** - Track hallucinations and reasoning quality

Implementation Plan

Question Dataset

- Web scraper for USMLE Step 1 questions
- Image capture and storage
- Structured data formatting

Model Integration

- API key collection and management
- Standard interface development
- Error handling implementation

Implementation Plan (cont.)

Evaluation System

- Response parsing algorithms
- Answer comparison logic
- Result logging system

Results Dashboard

- Accuracy visualization
- Detailed question logs
- Filtering and search capabilities
- Performance charts

User Stories

Core Requirements

- **Researcher:** Access standardized USMLE dataset for fair AI evaluation
- **Developer:** Integrate multiple AI providers through unified interface
- **Evaluator:** Automated answer checking for objective scoring at scale
- **Medical Educator:** Results dashboard for quick performance comparison
- **Student:** Identify best-performing AI model for reliable study assistance

Stretch User Stories

Advanced Requirements

- **Researcher:** Test models with/without images to measure visual reasoning
- **Developer:** Test prompt variations to evaluate model consistency
- **Clinician:** Test with additional lab values to assess medical reasoning depth

Next Steps

Week 2 Priorities

1. Set up project structure and environment
2. Begin USMLE question scraper development
3. Research and test AI API integrations
4. Design database schema for results storage

Goal: Have functional question dataset and basic model integration ready for testing