Feb 10, 2025

بسم الله الرحمن الرحيم

Islamic GenAI Guild
LLM Evaluation

Initiative 3

Version	Date	Author/Editor	Summary of Changes	Approved By
0.1	Feb 10, 2025	Atif Kureishy	Initial Draft	

Overview

This initiative establishes an agentic framework using LangGraph and LangSmith to evaluate Large Language Models (LLMs) on Islamic knowledge and ethics. The framework will be implemented as a multi-agent system, and the findings will be displayed on a dynamic website for public analysis.

Objectives

1. Develop a multi-agent system using LangGraph and LangSmith to automate LLM evaluations.
2. Establish a structured evaluation framework covering accuracy, ethical alignment, and citation correctness.
3. Implement an AI tool grading system based on evaluation criteria.
4. Develop a leaderboard ranking AI models based on their performance.
5. Create a WordPress-based website to showcase evaluation results and insights, built by an automated agent.

Technical Components

Multi-Agent System Nodes (Hikmah AI Framework)

Agent 1A: Ilm Extractor (Dataset Utilization)

- Function: Extracts and formats data from existing Islamic datasets.
- Implementation:
    - Define a node in LangGraph for dataset extraction and preprocessing.
    - Use Python scripts to process structured datasets for evaluation.
- Output: JSON-formatted structured queries from existing sources.

Agent 1B: Tafakkur Synthesizer (Synthetic Data Generator)

- Function: Generates new synthetic datasets for model evaluation.
- Implementation:
    - Define a node in LangGraph for generating synthetic Islamic Q&A datasets.
    - Use generative techniques to create diverse, unbiased evaluation sets.
- Output: JSON-formatted synthetic datasets for evaluation.

Agent 2: Adl Evaluator (Multi-LLM Evaluation & Response Scoring)

- Function: Evaluates LLM responses on:
    - Accuracy (based on authentic sources).
    - Ethical alignment (adherence to Islamic values).
    - Bias detection.
    - Source citation.
- Implementation:
    - Define nodes in LangGraph for each evaluation criterion.
    - Use LangSmith to track and compare model responses.
- Output: JSON-formatted evaluation results.

Agent 3: Mizan Ranker (Leaderboard & Reporting)

- Function: Aggregates evaluation scores, ranks models, and provides insights.
- Implementation:
    - Define a node in LangGraph to aggregate and rank AI performance metrics.
    - Use Pandas for data aggregation and reporting.
- Output: Ranked AI performance metrics and leaderboard data.

Agent 4: Bayan Publisher (Website Creation - WordPress)

- Function: Develops and maintains the WordPress-based website to display evaluation results dynamically.
- Implementation:
    - Define a node in LangGraph that automates WordPress site setup and content updates.
    - Use WordPress API to publish leaderboard rankings and evaluation reports.
    - Ensure mobile responsiveness and SEO optimization.
- Output: A fully functional, automated WordPress website presenting AI evaluation data.

Website Development

- Platform: WordPress (automated setup via an agent).
- Hosting: Hosting on wpengine.com as a sub-domain of mccsandiego.org.
- Features:
    - Dynamic leaderboard display.
    - Interactive visualization of AI evaluation results.
    - Periodic automated updates via LangGraph agents.

Timeline

Phase	Task	Timeline
1	Setup LangGraph & LangSmith	2 weeks
2	Implement Multi-Agent Nodes	4 weeks
3	Develop AI Evaluation Pipeline	3 weeks
4	Create WordPress Website	4 weeks
5	Testing & Optimization	3 weeks
6	Deployment & Public Release	2 weeks

Appendix: Defining Agents in LangGraph

1. Define Agents in LangGraph

- Create a StateGraph in LangGraph to model the workflow of each agent.
- Define nodes representing different functions or evaluations.
- Establish edges to determine the flow between nodes.

2. Utilize LangGraph Studio

- Connect the local agent to LangGraph Studio for visualization and debugging.
- Use the studio to test and refine agent workflows.

References:

- LangGraph Introduction
- Connecting a Local Agent to LangGraph Studio

This initiative aims to advance AI evaluation within an Islamic ethical framework, ensuring fairness, accuracy, and reliability in AI interactions with Islamic knowledge. May this effort be of benefit to the broader community.