FBS-2844: Propose an evaluation framework for an LLM chatbot.

Overview

Model Shortlisted

1. GPT-NeoX 20B (MIT)
2. GPT-J 6B (Apache 2.0)
3. OpenLLaMa 7B (Apache 2.0)
4. Falcon 40B (Apache 2.0)

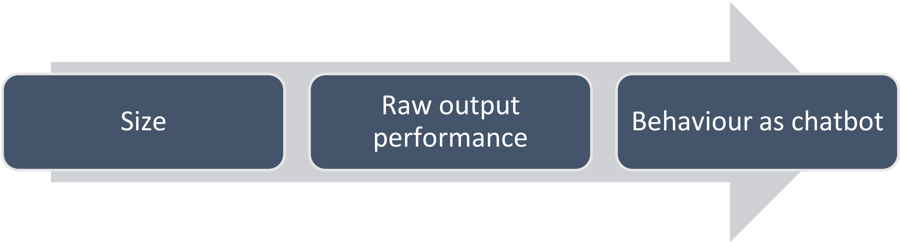
Criteria

- **Size of model**
  - Larger models requiring more powerful hardware and incurring higher costs.
  - Smaller models are more efficient to train, deploying them incurring lower costs.
  - Choosing the smallest possible model that meets our use case, we can optimize our costs and resources.
- **Performance – I: Use case specific**
  - Trained on Technical dataset for better recommendation.
  - RedPajama, Common crawl, C4, GitHub, Book, ArXiv, Wikipedia, Stack Exchange
- **Performance – II: Behavior as Chatbot**
  - How to Evaluate the Quality of LLM-based Chatbots | by Matt Ambrogi | Towards AI
  - **Chatbot evaluation at a high level**
    - What is a good response?
      - **Strategy – forms an intuitive opinion** done by asking a lot of questions and getting a feel for whether your bot's responses tend to be good or not.
    - How is qualitative evaluation done?
      - **Strategy – thumbs integration:** thumbs up / thumbs down next to a response that allows users to report whether the answer was helpful or not.

Insights

- For our use case **MIT and Apache 2.0 license** is relevant.
- **Open Flamingo** has 9B parameter size and dataset is large multimodal dataset c4 – document, images interleaved.
- **GPT-NeoX and GPT-J 20B** and 6B has parameter size and dataset is Pile An 800GB Dataset of Diverse Text for Language Modeling derive from academic or professional sources.
- **OpenLLaMa** has 7B parameter size and dataset is RedPajama dataset contains Common crawl, C4, GitHub, Book, ArXiv, Wikipedia, Stack Exchange
- **StableLM** has 3B parameters size and dataset is higher quality data sources and mixtures; specifically, the use of RefinedWeb (is built through stringent filtering and large-scale deduplication of CommonCrawl) and C4 (is the processed version of Google's C4 dataset).
- **Dolly** has 12B parameter size and dataset is capability domains from the InstructGPT paper, including brainstorming, classification, closed QA, generation, information extraction, open QA and summarization. Dataset is licensed for commercial use.
- **Falcon** has 40B parameter size and dataset was collected from public crawls of the web to build the pretraining dataset of Falcon. Using dumps from CommonCrawl, after significant filtering then extended with a few curated sources such as research papers and conversations from social media.

Framework for Evaluation



1. **Size of model**
   Models are already selected on basis of size in mind.

   Model Shortlisted

   1. GPT-NeoX 20B (MIT)
   2. GPT-J 6B (Apache 2.0)
   3. OpenLLaMa 7B (Apache 2.0)
   4. Falcon 40B (Apache 2.0)

2. **Performance – I: Use case specific**
   How the model performance without any filter?
   How raw model output relevant to ford specific software without any tweaking?
   Quantify the output to our use case.

   Approach: Creating Custom Score for comparison.

   1. Given list of known ford software
   2. Standard Prompt is given to all Models.
   3. Output generated is used to check similarity with given list of ford specific software
   4. Get cosine similarity scores for the output
   5. Repeat this for 10 queries and take an average to determine the score for evaluation.
   6. Compare models score for their performance without any filter.
   7. Select the best model.

3. **Performance – II: Behavior as Chatbot**
   How the model behavior?

   Approach: Chatbot evaluation at a high level

   1. What is a good response?
      Strategy – forms an intuitive opinion done by asking a lot of questions and getting a feel for whether your bot's responses tend to be good or not.
   2. How is qualitative evaluation done?
      Strategy – thumbs integration: thumbs up / thumbs down next to a response that allows users to report whether the answer was helpful or not.

Prompts for citizen developer.

Data Visualization:

- "I need assistance in creating a data visualization dashboard for my sales data. How can I get started?"
- **"What are the best tools and libraries to create interactive data visualizations from CSV files?"**
- "Can you help me design a chart that effectively displays monthly revenue trends?"

Data Analysis:

- "I have a dataset of customer reviews. How can I perform sentiment analysis on it?"
- "What statistical methods can I use to analyze user engagement data for my website?"
- "How can I clean and preprocess data from multiple sources before performing an analysis?"

Development of a Project:

- **"I want to build a mobile app for tracking fitness goals. What tech stack should I consider for this project?"**
- "What are the best practices for developing a web-based inventory management system?"
- "Can you provide a step-by-step guide for setting up a version control system for my software project?"

Web Development Basics:

- **"I'm new to web development. What are the fundamental technologies and languages I should explore?"**
- "How can I create a simple webpage with HTML and CSS?"
- "What are the key differences between frontend and backend development?"

Frontend Development:

- "I want to build a responsive and user-friendly website. Can you recommend a frontend framework or library?"
- "How can I add interactivity to my website using JavaScript?"
- "What are some best practices for optimizing the performance of my web application?"

Backend Development:

- **"What are the popular backend programming languages and frameworks for web development?"**
- "How can I set up a database for my web application, and what type of database should I choose?"
- "What are RESTful APIs, and how can I create one for my web service?"

Web Development Tools:

- **"What code editor or integrated development environment (IDE) do you recommend for web development?"**
- "How can I deploy my website or web application to a hosting platform?"
- "What are the steps to secure my web application from common vulnerabilities?"