Report written in Markdown - website is live at https://pndang.com/GenAI_Capstone_ProjectOne/

Data Science Capstone Project Report

Result-replication instructions:

1. Clone this repository to local development environment.
2. Navigate to the working directory and install the dependencies in **[requirements.txt]**.
3. Once done, run the command **[streamlit run app.py]** to locally serve and use the streamlit app LLM translator.
4. Other project works can be found in the **[.ipynb]** files (Jupyter notebooks), which are more "relaxed" regarding dependencies, simply run the notebooks (with **[!pip install]** commands) to reproduce the notebooks' results.
5. Enjoy being a Guardian of the Generative Realm!

Author: Phu Dang

Capstone Team: Deloitte X HDSI, UC San Diego

Guardians 🦸 of the Generative Realm 👾: Implementing and Benchmarking Chatbot Guardrails

Abstract

The term **AI** — particularly Generative AI — has increasingly becoming an integral part of individuals' lives and enterprises, with a strong focus on improvements, such as efficiency, automation, cost reduction, and companionship — however — seldom are the conversations about risk-mitigation and the potential harms of GenAI, which are key to building **trust** for the effective and ethical adoption of GenAI at both the indiviudal- and enterprise-level. The work of this survey project aims to implement a wide variety of LLM guardrail frameworks, balancing both breadth and depth, to document the pros and cons of each framework (qualitative) and benchmark the performance for each framework (quantitative) for analysis and comparison. Key measurements are implementation complexity, ease-of-use, robustness, fault tolerance, security, latency, and prompt-flag accuracy. [PLACE HOLDER FOR FUTURE RESULTS AND FINDINGS].

Introduction

Background

In "The great acceleration: CIO perspectives on generative AI" by the MIT Tech Review and Databricks, Schaefer—Chief Health Informatics Officer at Kansas City VA Medical Center—stated **"trust"** as the Key to effectively adopting generative AI. The healthcare industry is a great case study for involving a multitude of high-stake processes and parties, such as the use of machine learning models to predict protein structures, assist drug discovery, and track the progression of outbreaks, to chatbots helping front-line staff by transcribing medical notes and answering patients' questions. The same understanding of such broad applications can be applied to other industries, where enterprises are less efficient with scattered data sources — this struggle presents an immense opportunity for generative AI to unify such scattered information into productive uses through large language models. Realizing this opportunity requires its effective and ethical implementation to protect data and network integrity, safety, and privacy across systems for the collective benefits of those affected most by generative AI.

Literature Review

Many prominent works have been done to document, understand, and disseminate the risks and threats of generative AI, notably is the OWASP Top 10 security risks for large language model (LLM) applications, which includes prompt injection, insecure output handling, sensitive information leakage, excessive agency, to model theft. Knowing these preeminent threats allows developers and users of LLM-based chatbots to navigate development environments, workflows, usage, and system architectures with proactive risk-mitigation techniques to address these concerns.

One popular LLM guardrail framework has been provided by Guardrails AI that focuses on input and output handling to ensure harmful and sensitive information are not leaked. Guardrails AI implements many state-of-the-art programming techniques to minimize latency while maintaining stable accuracy to improve user experience, such as parallel programming, which has proven especially useful for chatbots peforming data-intensive processes, such as Retrieval-Augmented Generation (RAG). Special emphases are also placed on robustness and controllabiltiy of the development environment, with documented decision-making processes for the most stable and reliable product. (Shared in guest lecture from Zayd Simjee, Co-Founder of Guardrails AI)

[PLACEHOLDER FOR DISCUSSION OF NEMO GUARDRAILS, OPENAI COOKBOOK GUARDRAILS, ETC.]

Data Processes Description

The points below capture the key focuses of this project. Since the focus is on implementing and benchmarking LLM guardrails, little data will be used throughout the project, unless for occasional speed tests to document the latency of different guardrail frameworks at performing data-intensive techniques, such as Retrieval-Augmented Generation.

1. **Multimodal Data Connectivity** - Automating the process of gathering, cleaning, and transforming raw data from various sources by mimicking a real-world enterprise scenario to address the problem involving scattered information sources.
2. **Robust Prompt Monitoring** - Deploying robust, fault rolerant, and mindful input and output prompt-checking techniques (guardrails) to ensure harmful, sensitive, and riskful information are not leaked to malicious actors, nor communicated to the users, not allowed as inputs to the model.
3. **Ethical User Considerations** - Actively plan and orchestrate scenarios of attack to improve the future applicability and effectiveness of LLM guardrails in enterprise applications. High attentiveness will be given to define and design user flows, allowing the guardrail frameworks to address a wide-ranging scope of situations users may face interacting with LLM-based chatbots.

More in progress!