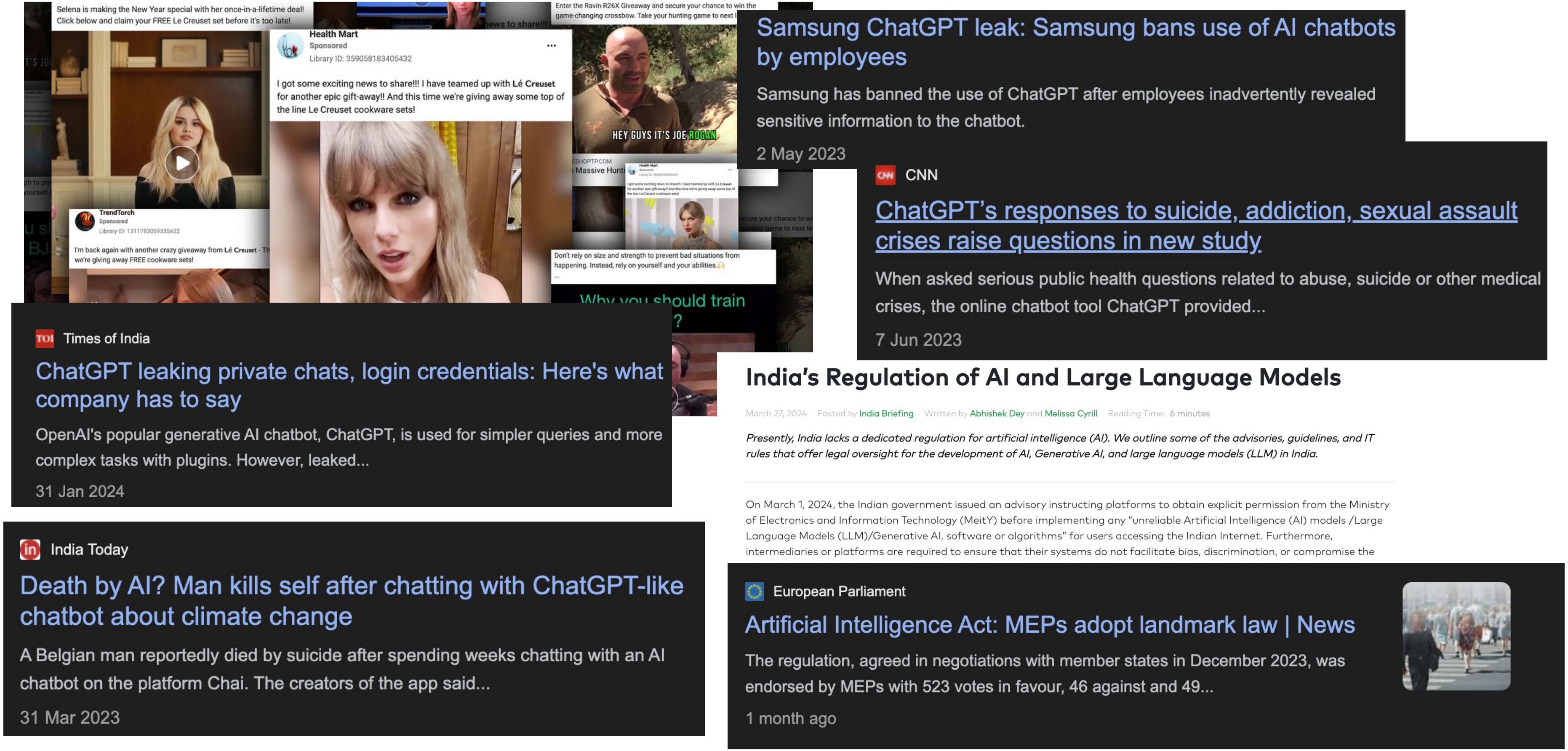
EthiSecure.AI : Keeping AI Chats Secure and Ethical

Team : F16Falcon

Anannyo Dey Debasmit Roy Kabir Raj Singh Aditya Ganguly

Recent events related to Generative AI



Selena is making the New Year special with her once-in-a-lifetime deal! Click below and claim your FREE Le Creuset set before it's too late!

Health Mart
Sponsored
Library ID: 359058183405432

I got some exciting news to share!!! I have teamed up with Lé Creuset for another epic gift-away!! And this time we're giving away some top of the line Le Creuset cookware sets!

TrendTorch
Sponsored
Library ID: 1311782059535622

I'm back again with another crazy giveaway from Lé Creuset - Th we're giving away FREE cookware sets!

Enter the Ravin R26X Giveaway and secure your chance to win the game-changing crossbow. Take your hunting game to next l

Don't rely on size and strength to prevent bad situations from happening. Instead, rely on yourself and your abilities.🙌

Samsung ChatGPT leak: Samsung bans use of AI chatbots by employees

Samsung has banned the use of ChatGPT after employees inadvertently revealed sensitive information to the chatbot.

2 May 2023

CNN

ChatGPT's responses to suicide, addiction, sexual assault crises raise questions in new study

When asked serious public health questions related to abuse, suicide or other medical crises, the online chatbot tool ChatGPT provided...

7 Jun 2023

Times of India

ChatGPT leaking private chats, login credentials: Here's what company has to say

OpenAI's popular generative AI chatbot, ChatGPT, is used for simpler queries and more complex tasks with plugins. However, leaked...

31 Jan 2024

India's Regulation of AI and Large Language Models

March 27, 2024 Posted by India Briefing Written by Abhishek Dey and Melissa Cyrill Reading Time: 6 minutes

Presently, India lacks a dedicated regulation for artificial intelligence (AI). We outline some of the advisories, guidelines, and IT rules that offer legal oversight for the development of AI, Generative AI, and large language models (LLM) in India.

On March 1, 2024, the Indian government issued an advisory instructing platforms to obtain explicit permission from the Ministry of Electronics and Information Technology (MeitY) before implementing any "unreliable Artificial Intelligence (AI) models /Large Language Models (LLM)/Generative AI, software or algorithms" for users accessing the Indian Internet. Furthermore, intermediaries or platforms are required to ensure that their systems do not facilitate bias, discrimination, or compromise the

India Today

Death by AI? Man kills self after chatting with ChatGPT-like chatbot about climate change

A Belgian man reportedly died by suicide after spending weeks chatting with an AI chatbot on the platform Chai. The creators of the app said...

31 Mar 2023

European Parliament

Artificial Intelligence Act: MEPs adopt landmark law | News

The regulation, agreed in negotiations with member states in December 2023, was endorsed by MEPs with 523 votes in favour, 46 against and 49...

1 month ago

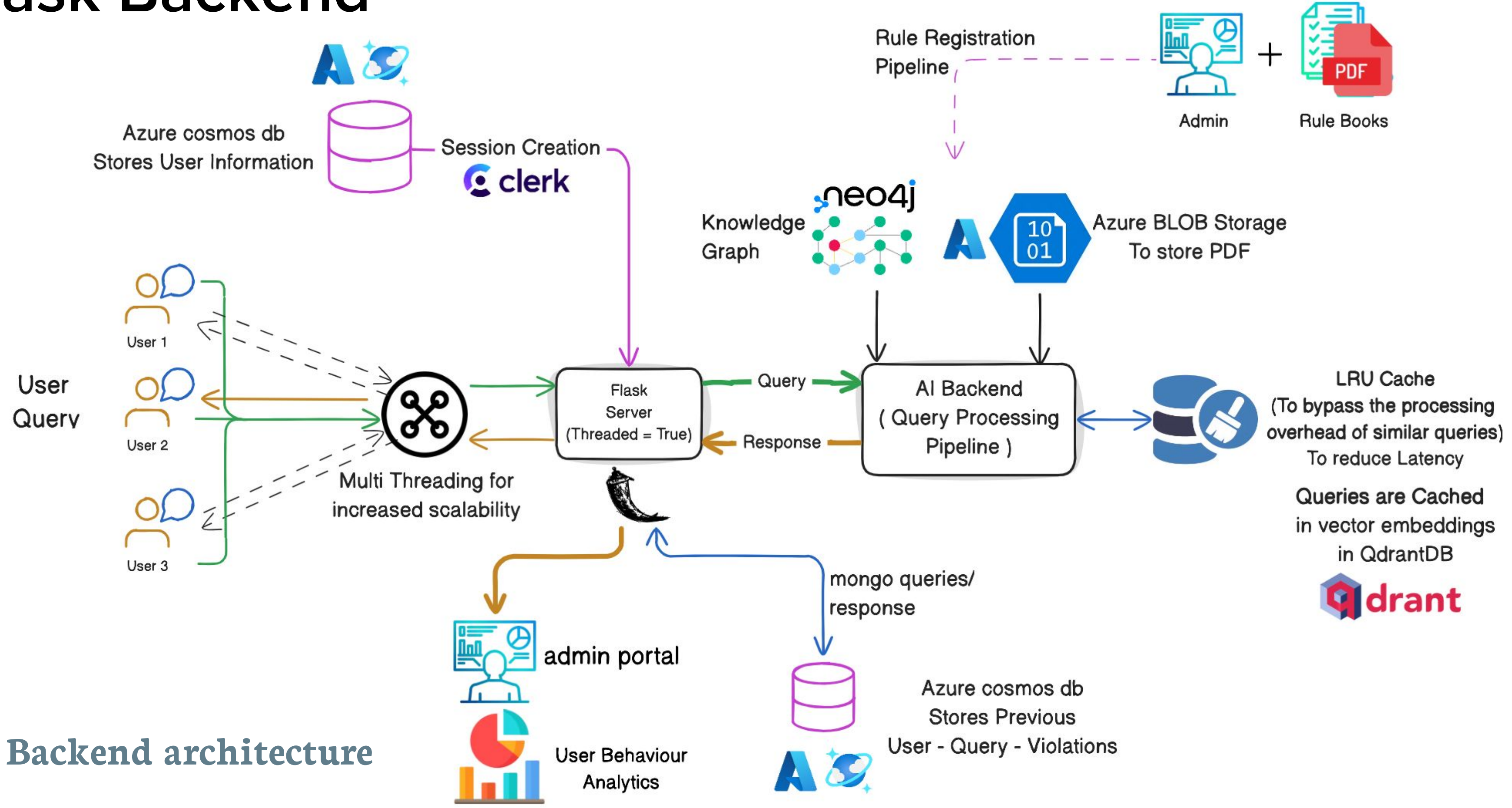
Approach

- Usage of **MS Azure CosmosDB** for storing user info and violations
- Implementation of **multithreading** in Flask for increased scalability
- Integration of **LRU cache** to bypass the processing overhead for similar queries, thereby reducing latency
- Development of LLM-powered Rule keyword extractor for creating **Rule Knowledge graph** (KG) from PDFs uploaded by admin for **complex RAG.** It can be seen as an **alternative of expensive fine-tuning.**
- **Azure Blob storage** to store Admin's Rule PDFs and **neo4j** to store the **KG**.
- **Semantic router** to retrieve relevant **subgraphs** based on user queries/LLM responses, followed by a **query pipeline** to identify violations (if any).
- User authentication using **Clerk** integrated with **Azure CosmosDB.**
- Deployment of entire solution on **Vercel (frontend) and Pythonanywhere (backend) and MS Azure (database and storage)**

Features

- Periodically updated Rule **Knowledge graph**, whenever admin uploads new PDFs
- Display of **rule violations and reasons** levels on admin dashboard
- Assignment of **Risk score based on number of violations and its risk levels**
- User behaviour pattern analysis : **Category wise Rule Violation Count per day/week**
- **MS Azure** integration performed utilizing **Cosmos DB** containers and **Blob storage**

Flask Backend



Azure cosmos db
Stores User Information

Session Creation
clerk

Rule Registration
Pipeline

Admin

Rule Books

Knowledge
Graph

neo4j

Azure BLOB Storage
To store PDF

User
Query

User 1

User 2

User 3

Multi Threading for
increased scalability

Flask
Server
(Threaded = True)

Query

Response

AI Backend
(Query Processing
Pipeline)

LRU Cache
(To bypass the processing
overhead of similar queries)
To reduce Latency

Queries are Cached
in vector embeddings
in QdrantDB

qdrant

mongo queries/
response

admin portal

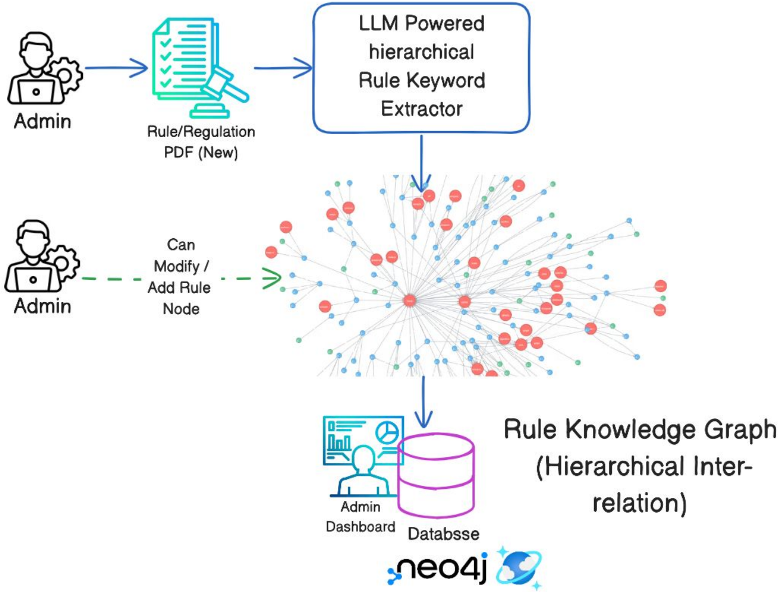
User Behaviour
Analytics

Azure cosmos db
Stores Previous
User - Query - Violations

Backend architecture

High Level Design AI Backend

Rule Registration Pipeline



Admin

Rule/Regulation PDF (New)

LLM Powered hierarchical Rule Keyword Extractor

Admin

Can Modify / Add Rule Node

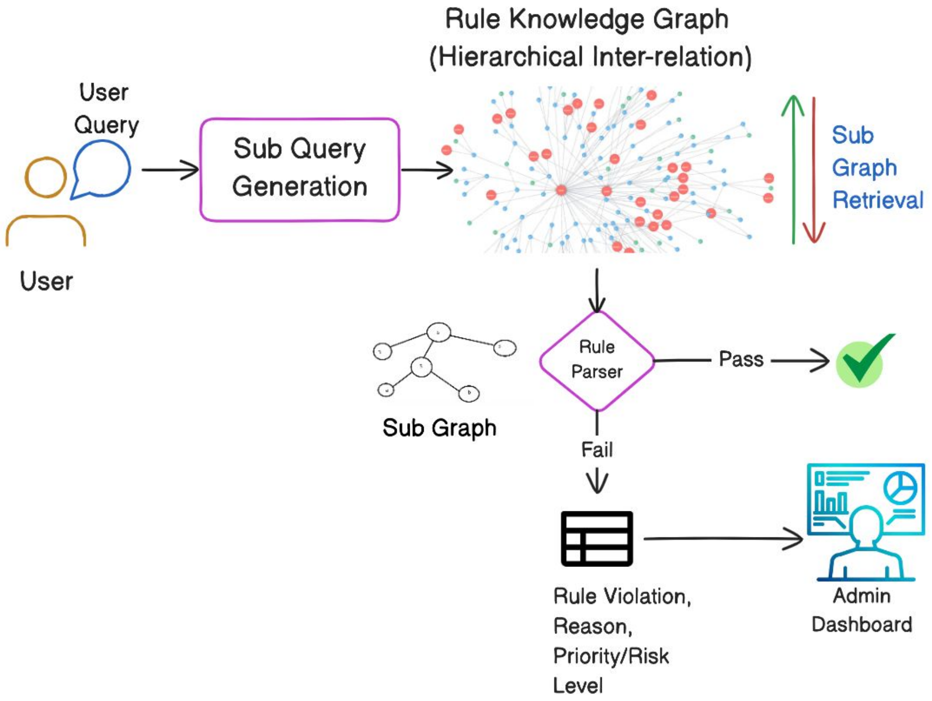
Admin Dashboard

Databsse

neo4j

Rule Knowledge Graph (Hierarchical Inter-relation)

Query Processing Pipeline



Rule Knowledge Graph (Hierarchical Inter-relation)

User Query

User

Sub Query Generation

Sub Graph Retrieval

Sub Graph

Rule Parser

Pass

Fail

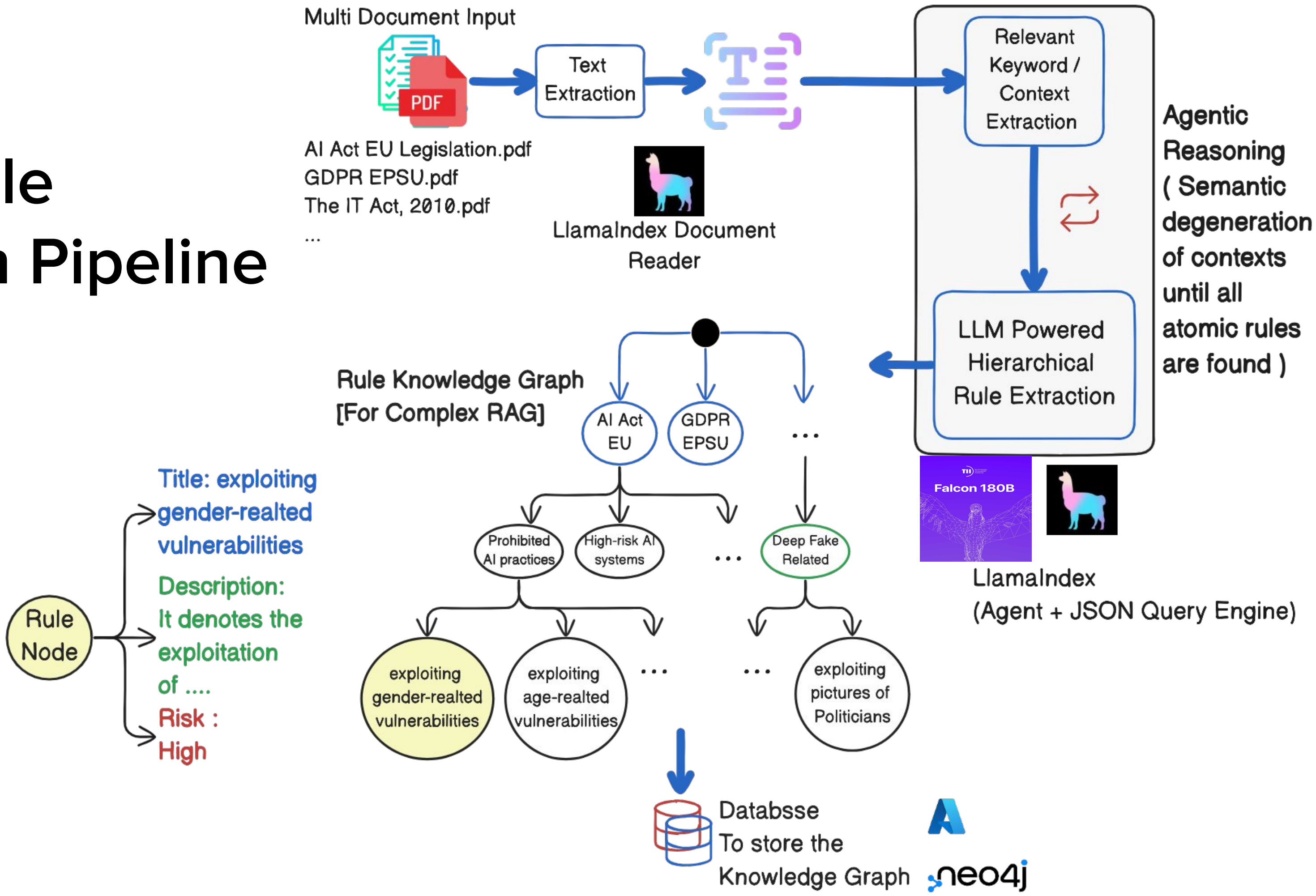
Rule Violation, Reason, Priority/Risk Level

Admin Dashboard

Rule Registration Pipeline

1. The admin sends PDFs containing Rule to the system, which is stored on **Azure CosmosDB** then passed to the **Llama Index document reader** to extract the text.
2. The entire PDF text is decomposed into multiple contexts, which are then further broken down into more atomic components. Utilizing the Llama Index powered agentic reasoning pipeline, we identify and decompose relevant contexts until no additional rules can be extracted.
3. Similar rule phrases are grouped to form a **knowledge graph**, delineating the **hierarchical relations** among several rules. This will be further used to parse **complex RAG queries**. Risk levels are assigned for each terminal nodes as mentioned in the pdf document. **Knowledge graph** is stored in **neo4j graph-db.**
4. **External knowledge source** also integrated to check rules related to recent topics.
5. The admin can also **modify** the AI generated tree structure from Admin dashboard if necessary, by **adding, deleting or editing items**

Detailed Rule Registration Pipeline



Multi Document Input

AI Act EU Legislation.pdf
GDPR EPSU.pdf
The IT Act, 2010.pdf
...

Text Extraction

LlamaIndex Document Reader

Relevant Keyword / Context Extraction

LLM Powered Hierarchical Rule Extraction

Agentic Reasoning (Semantic degeneration of contexts until all atomic rules are found)

LlamaIndex (Agent + JSON Query Engine)

Rule Knowledge Graph [For Complex RAG]

AI Act EU
GDPR EPSU
...

Prohibited AI practices
High-risk AI systems
...
Deep Fake Related

exploiting gender-realted vulnerabilities
exploiting age-realted vulnerabilities
...
...
exploiting pictures of Politicians

Rule Node

Title: exploiting gender-realted vulnerabilities

Description: It denotes the exploitation of ....

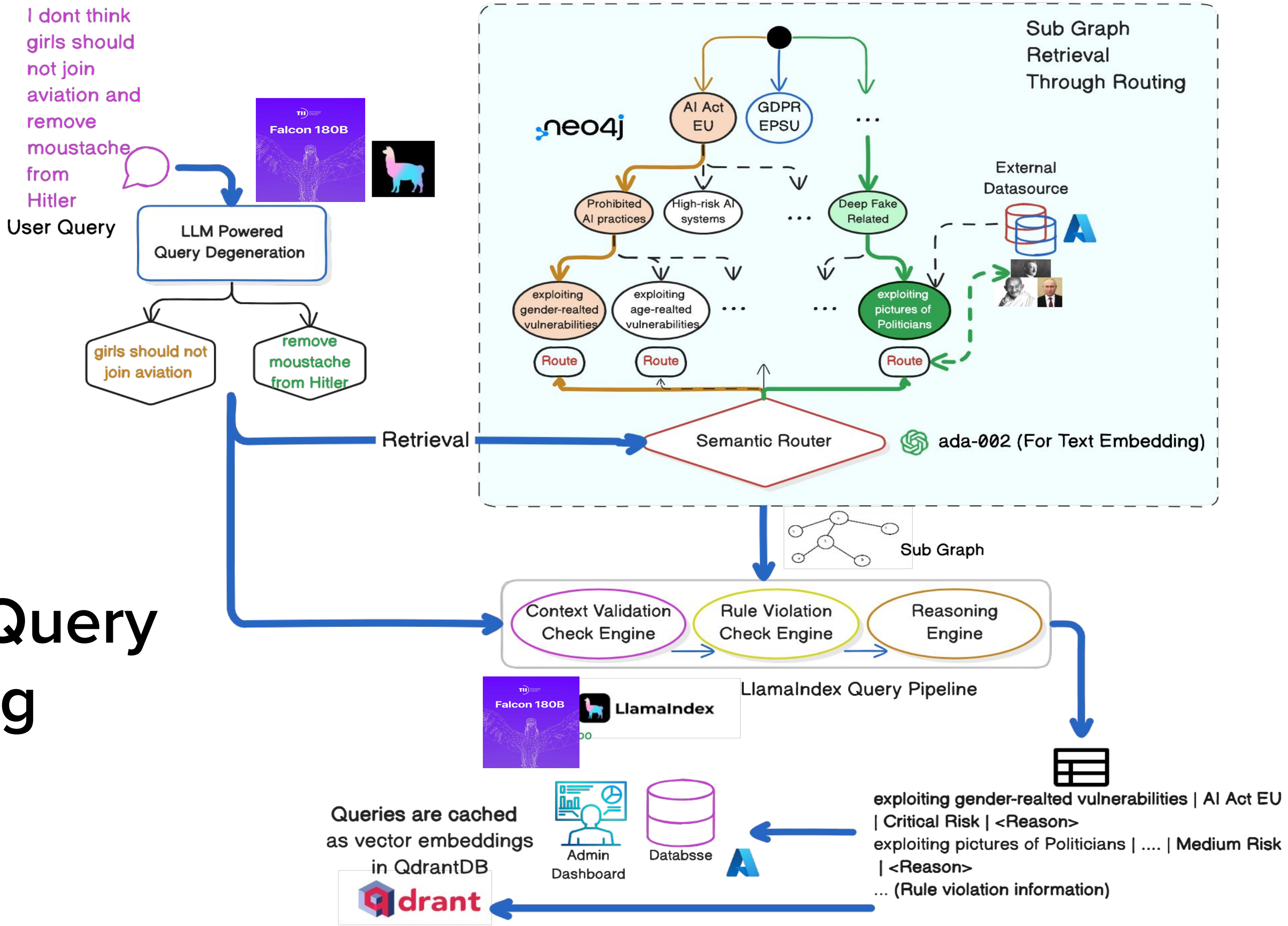
Risk : High

Databsse To store the Knowledge Graph

Query Processing Pipeline

1. Firstly, user query is degenerated into subqueries using **Llama Index JSON Query Engine**.
2. These subqueries are then routed to the existing knowledge graph through a **semantic router**, using **text-embedding-ada-002** for semantic similarity calculation. Subsequently, all relevant paths in the form of **subgraphs** are retrieved.
3. The **subqueries** along with the relevant **subgraph** are ingested into a **LlamIndex Query Pipeline** where **three sequential LLM engines** operates:
    i. **Context Validation Check Engine** validates whether **subqueries** and **subgraph** belong to a same contexts
    ii. **Rule Violation Check Engine** then check whether the **subqueries** violate any rule from subgraph
    iii. **Reasoning Engine** simultaneously states the **reason** for violation if any for each subquery.
4. All possible violations are listed along with calculated risk count. These are stored on **MS Azure CosmosDB containers.**
5. Queries are **cached** as vector embeddings in **Qdrant Vector DB** along with the response to reduce the processing overhead of very similar queries.

Detailed Query Processing Pipeline



I dont think girls should not join aviation and remove moustache from Hitler

User Query

Falcon 180B

LLM Powered Query Degeneration

girls should not join aviation

remove moustache from Hitler

Retrieval

Sub Graph Retrieval Through Routing

neo4j

AI Act EU

GDPR EPSU

. . .

Prohibited AI practices

High-risk AI systems

. . .

Deep Fake Related

External Datasource

exploiting gender-realted vulnerabilities

exploiting age-realted vulnerabilities

. . .

. . .

exploiting pictures of Politicians

Route

Route

Route

Semantic Router

ada-002 (For Text Embedding)

Sub Graph

Context Validation Check Engine

Rule Violation Check Engine

Reasoning Engine

Falcon 180B

LlamaIndex

LlamaIndex Query Pipeline

exploiting gender-realted vulnerabilities | AI Act EU | Critical Risk | <Reason>

exploiting pictures of Politicians | .... | Medium Risk | <Reason>

... (Rule violation information)

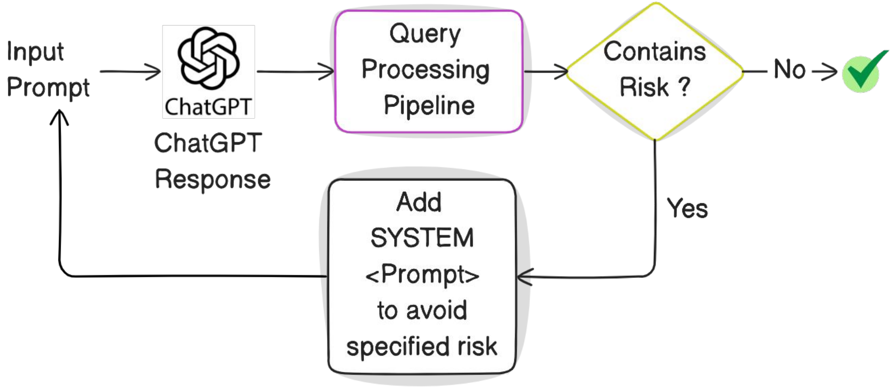
Queries are cached as vector embeddings in QdrantDB

Admin Dashboard

Databsse

drant

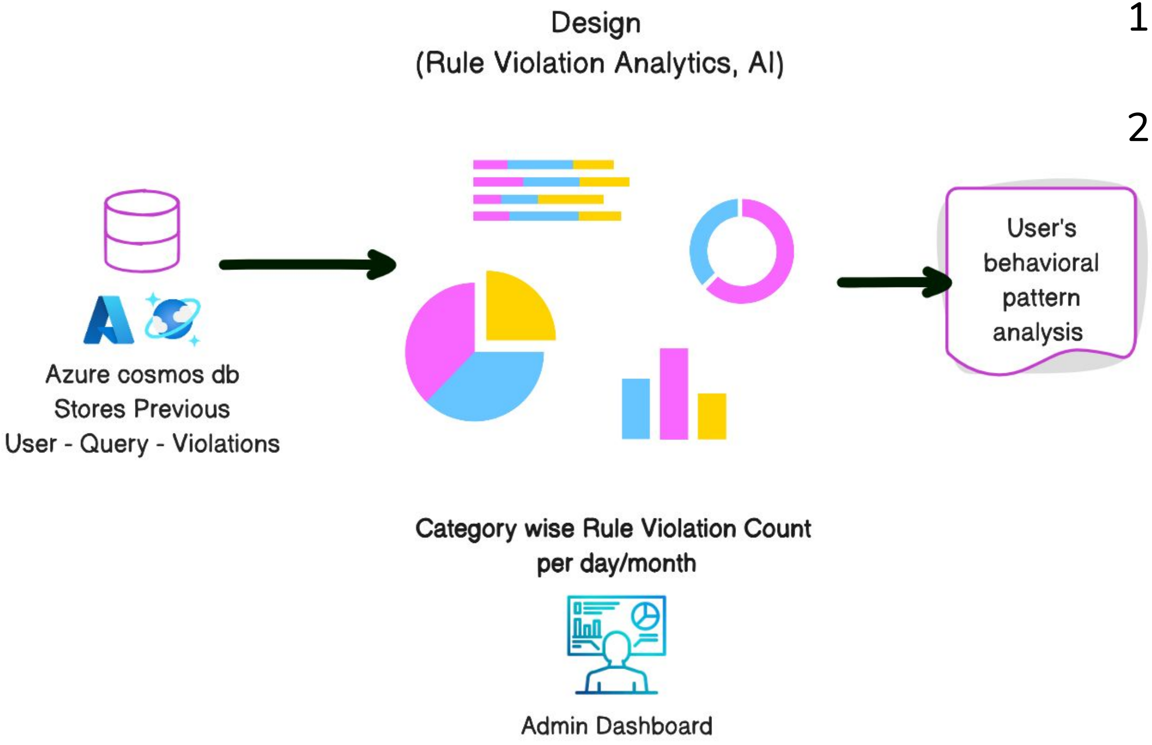
Mitigating risky ChatGPT response


Input Prompt
ChatGPT
ChatGPT Response
Query Processing Pipeline
Contains Risk ?
No
Yes
Add SYSTEM <Prompt> to avoid specified risk

Risk Scoring Algorithm

We assign risk scores to every violation in the form of user query and LLM response. This is based on risk level categorization of the **EU AI act** - Critical, High, Medium & Minimal, where each level has a **defined score** and based on the number and nature of violations in the query/response we calculate the **Risk score**. We show risk levels and their categories in the admin analytics panel. **Risk thresholding** is performed on the scores calculated.

Analysis and Visualization



1. All violations for each user query are stored in **Azure Cosmos DB**
2. These data are displayed on **Admin Dashboard**, in form of charts-
   a. Line and Bar charts: **No. of violations** within **custom time-frame for each risk level**
   b. Stacked Chart: **Compare no. of Rule violations** per **high-level category** between 2 dates
   c. Pie chart: **Compares %** of **violations** and **safe queries**
   d. User-wise violation analytics: **Top5 violations per user**

Challenges and Solution

1. LLM Hallucination: We mitigate this by using **low temperature LLMs** and in-context RAG based search in our Rule parse tree (Knowledge Graph, KG). As we fragment the query into smaller phrases and then pass it through the router, we can avoid dependency on high-temperature language model (LLM).

2. Improving explainability: Parsing user queries and LLM responses using Rule parse tree KG and semantic router makes it **explainable both to the admin and end user**, as to **exactly which violations have occurred and their severity levels.**

3. Improving scalability: We support multiple users and LLM sessions simultaneously using **multithreading**. Admin has the privilege to **add more rule nodes** to the KG making it more scalable.

4. Reducing latency: We have taken several measure like using **LRU cache, mitigating the usage of high temperature LLM** etc. to reduce the response time and make the entire process real-time, even while handling multiple clients.

Tech Stack

Python

ReactJS

Chart.js

Flask

Clerk

Redux

OpenAI

LangChain

Qdrant Vector Store

neo4j

LlamaIndex

Azure

Cosmos DB

Blob Storage

Vercel

pythonanywhere
by ANACONDA

Models

1. LLM Model: Falcon-180b-chat (Temperature 0.0)
2. Dense Embedding Model : openai/text-embedding-ada-002

Future Scopes:

1. Integration of GPT4 models which can analyze multimodal queries for potential risks.
2. To safeguard responses based on publicly available code of conduct beyond the provided rulebooks.

Thank You

Link : https://ethicheck-ai-lemon.vercel.app/
Code: https://github.com/orgs/falcon-proj/repositories