AI Against Crime

Integrating Retrieval Augmented Generation with Contrastive Language-Image Pretraining in Criminal Network Analysis

AUTHOR:
Francisco Manuel Trujillo Montenegro

INTRODUCTION

This research introduces an advanced solution aimed at boosting the fight against high-risk criminal networks by automating information flow through the integration of Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) and Contrastive Language-Image Pretraining (CLIP). Our system adeptly processes and refines large-scale data from initial inputs to precise, actionable outputs. The collaboration between RAG and CLIP enables an enriched contextual understanding and enhances accuracy in analyzing multimodal data and producing reports that are easily comprehensible to investigators.

Focusing on its application in the law enforcement and security sectors, this innovative approach illustrates how AI can significantly enhance the capabilities of agencies in tracking and disrupting criminal activities.

The poster details the system's architecture, demonstrates its effectiveness in practical scenarios, and discusses the potential transformative impacts of employing AI agents in a data-driven orchestration within the context of combating high-risk criminal networks in an isolated environment.

OBJECTIVE

Hands-On AI Implementation: Design and deploy a proof-of-concept AI system using LLMs with RAG and CLIP.

Improve Data Analysis Accuracy and Efficiency: Increase precision and processing speed of multimodal data to support quicker and more informed real-time decision-making in criminal investigations.

Automate Reporting: Autonomously collects, analyzes, and generates detailed and accurate reports for use in law enforcement operations.

METHODOLOGY

A complex deployment setup using Docker containers has been implemented to manage five models: Mistral, Phi3, Whisper, ViT-B/32, and Llama3. These models are supplemented with the open-source tool Haystack to develop dynamic data retrieval pipelines.

A wide range of multimodal data is collected, including text documents, images, videos, and audio files from diverse sources. Data orchestration is utilized to automate the data flow, ensuring efficient preprocessing. This preprocessing includes data cleansing, normalization, and augmentation, leveraging AI agents.

RESULTS

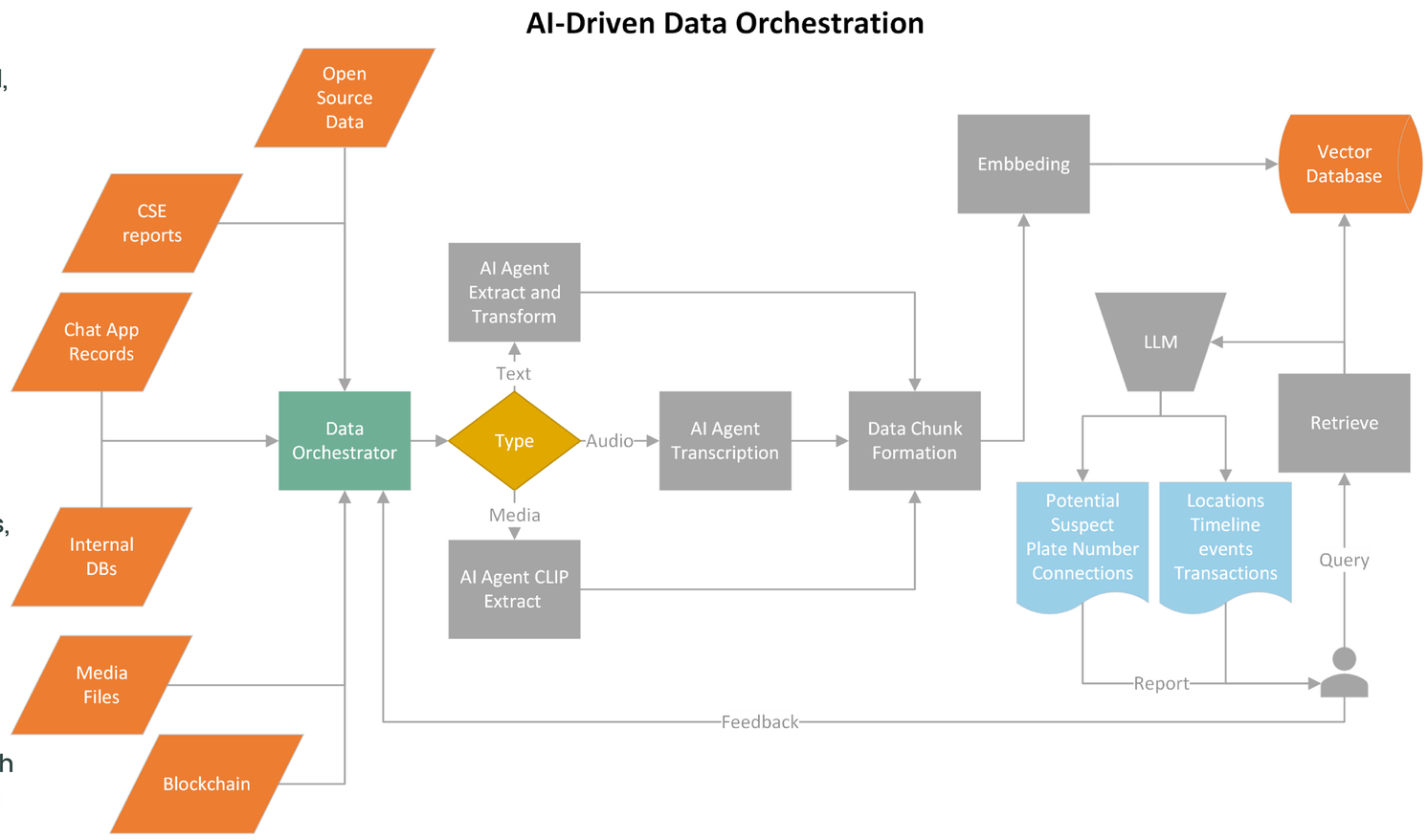
Enhanced Data Processing Efficiency: The integration of Docker with support to several LLMs has significantly increased the speed and accuracy of multimodal data analysis. This enhancement has been particularly effective in facilitating a better understanding of complex visual and audio evidence.

Improved Real-Time Intelligence: Data automation, combined with AI models, has enabled the real-time detection of criminal activities dispersed across a wide range of sources, thereby generating actionable multilingual intelligence.

ANALYSIS

The system's ability to operate offline is crucial, especially in secure law enforcement environments where internet access may be limited. It effectively manages a wide array of input data types, including images of people, vehicles, objects, various text formats in multiple languages, chats, internal databases entries, data breach records, location data, audios, and videos. This comprehensive data intake is critical for constructing a thorough understanding of criminal activities and networks.

The system consolidates its processed outputs, using the scalability of Docker containers, into a single portal, which is a significant operational advantage. This portal allows field agents and decision-makers to access synthesized intelligence reports, analytics, and actionable insights from one centralized location. The user can quickly navigate through complex datasets and retrieve the information they need without specialized technical skills.



The Data Orchestrator utilizes AI agents to extract, transform, and store diverse inputs into an organized structure, subsequently preparing the data to feed the LLM.

CONCLUSION

The system's robust offline capability, combined with its sophisticated AI-driven data processing and user-friendly portal, significantly boosts operational efficiency and effectiveness in combating high-risk criminal networks. Its comprehensive approach to handling and analyzing diverse data types ensures that law enforcement agencies have the actionable intelligence they need to respond swiftly and effectively to criminal threats.

LIMITATIONS

GPU Dependency: These models require significant resources, increasing costs and limiting scalability in resource-constrained environments.

LLM Hallucinations: LLMs can generate incorrect or irrelevant information, risking misinformed decisions in law enforcement.

Data Volume Management: Multimodal data demands advanced strategies to maintain system performance.

RELATED LITERATURE

RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. (2023). Available at: https://arxiv.org/abs/2310.04408 (Accessed: 14 May 2024).

Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. (2024). Available at: https://arxiv.org/abs/2403.14403 (Accessed: 14 May 2024).

Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. (2022). Available at: https://arxiv.org/abs/2109.02748 (Accessed: 14 May 2024).

Can Whisper Perform Speech-Based In-Context Learning? (2024). Available at: https://ieeexplore.ieee.org/abstract/document/10446502 (Accessed: 14 May 2024).

A Data Processing and Distribution System Based on Apache NiFi. (2023). Available at: https://www.mdpi.com/2304-6732/10/2/210 (Accessed: 14 May 2024).

Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. (2023). Available at: https://www.researchgate.net/publication/372855940_Llama_2_Early_Adopters'_Utilization_of_Meta's_New_Open-Source_Pretrained_Model (Accessed: 14 May 2024).