IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 April 2024, accepted 21 April 2024, date of publication 23 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3404862

RESEARCH ARTICLE

A Framework for LLM-Assisted Smart Policing System

PARIA SARZAEIM, QUSAY H. MAHMOUD, (Senior Member, IEEE), AND AKRAMUL AZIM, (Senior Member, IEEE)
Department of Electrical, Computer, and Software Engineering, Ontario Tech University, Oshawa, ON L1G 0C5, Canada

Corresponding author: Paria Sarzaeim (paria.sarzaeim@ontariotechu.net)

This work was supported in part by the Mitacs Accelerate collaborative research project with industry partner Mobile Innovations.

ABSTRACT In the face of rapidly increasing crime rates, the evolving complexity of crime data processing, and public safety challenges, the need for more advanced policing solutions has increased leading to the emergence of smart policing systems and predictive policing techniques. This urgency and shift toward smart policing incorporates artificial intelligence (AI), with a specific focus on machine learning (ML) as an essential tool for data analysis, pattern recognition, and proactive crime forecasting. Among these, the flexibility and power of AI techniques including large language models (LLMs), as a subset of generative AI, have increased the interest in applying them in real-world applications, such as financial, medical, legal, and agricultural applications. However, the abilities and possibilities of adopting LLMs in applications including crime prediction remain unexplored. This paper focuses on bridging this gap by developing a framework based on the transformative potential of BART, GPT-3, and GPT-4, three state-of-the-art LLMs, in the domain of smart policing, specifically, crime prediction. As a prototype, diverse methods such as zero-shot prompting, few-shot prompting, and fine-tuning are used to comprehensively assess the performance of these models in crime prediction based on state-of-the-art datasets from two major cities: San Francisco and Los Angeles. The main objective is to illuminate the adaptability of LLMs and their capacity to revolutionize crime analysis practices. Additionally, a comparative analysis of the aforementioned methods on the GPT series model and BART with ML techniques is provided which shows that the GPT models are more suitable than the traditional ML models for crime classification in most experimental scenarios.

INDEX TERMS Crime prediction, fine-tuning, few-shot prompting, large language models, LLM, zero-shot prompting.

I. INTRODUCTION

The rise in urbanization and its resultant increase in urban populations, cities face new challenges. Following this evolution, the increase in reported criminal incidents, accompanied by the growing amount of crime data, and the multiple challenges they pose to societies and public health and safety, have led to a growing demand for effective crime forecasting and prevention measures. Moreover, the main priority of police departments is to prevent crime by increasing the safety of cities. In this context, traditional solutions have proven insufficient for analyzing complex crime data and providing timely insights into potential incidents [1], [2]. Smart policing has emerged in this landscape in response to the urgent need for innovative solutions in law enforcement [3], [4], and it offers a framework for smart technologies powered by AI techniques such as data analysis and pattern recognition enabling authorities to identify emerging criminal patterns and trends effectively. These tools can potentially speed up data processing and analysis for law enforcement while mitigating the influence of human biases [5].

With the advancements in smart policing in recent years, most police departments use electronic systems for crime reporting, which have replaced traditional paper-based crime reports. Additionally, they commonly use AI tools for different purposes. Each of these technologies is used for

The associate editor coordinating the review of this manuscript and approving it for publication was Wojciech Sałabun.

various purposes including crime information extraction, traffic management, facial recognition, weapon detection, and monitoring criminal activity through surveillance cameras. [6], [7], [8] Furthermore, predictive policing has been introduced as a research subfield of smart policing, which involves using a range of technologies, such as crime documentation, predictive crime maps, advanced computer software, and artificial intelligence algorithms. Predictive policing enables police to use predictive analytics, forecast the occurrence of future crimes, and identify potential criminals and victims. Predictive policing leverages ML algorithms and statistical analysis methods to forecast criminal activities, including key details such as location, date, time, crime type, and potential victims, by analyzing both historical and real-time crime data [9]. The underlying idea of predictive policing is rooted in the theory that crimes do not occur randomly; instead, they follow patterns influenced by local environmental conditions and situational decision-making of potential victims [10].

LLMs are a class of generative AI models based on deep learning specifically designed for comprehending and generating human language. These models are characterized by their enormous size, often containing hundreds of millions or even billions of parameters, which allows them to perform a wide range of natural language understanding and text generation tasks [11]. For crime prediction tasks, an ML model can be trained on inputs to output labels from a fixed set of classes. However, recent advances in LLMs, such as BERT [12] and GPT-3 [11], have introduced a novel approach for classification known as prompting. In prompting, there is typically no need for further training; instead, the model input contains a task-specific text called a prompt. Prompt engineering involves creating, evaluating, and recommending prompts for natural language processing tasks, enabling professionals to use LLMs for tasks such as data annotation, classification, and question-answering. Another way to take advantage of powerful LLMs is to adapt them for specific tasks by fine-tuning them.

Despite their proven efficacy, using LLMs in smart policing, particularly for predictive policing applications, remains relatively unexplored. This paper addresses this gap by investigating the potential adaptability of LLMs within the domain of smart policing, with a specific focus on predictive policing using classification and prediction. The proposed framework extends the work in [13] and utilizes prompt engineering and fine-tuning methods of LLMs such as BART and GPT models. For the prototype, two state-of-the-art historical crime datasets from San Francisco and Los Angeles were used for crime prediction and classification tasks. In addition, a comparative analysis between LLMs and classical ML models in crime prediction on the same datasets is provided. The findings offer valuable insights into the practical implementation of LLMs in predictive policing. Through this research, we aim to shed light on the potential of LLMs to revolutionize crime analysis practices and enhance the efficacy of predictive policing techniques. To this end, the remainder of this paper is organized as follows. Section II provides an overview of related work on crime classification and prediction, along with background information on LLMs. Section III clarifies the prompting and fine-tuning methods used. Section IV explains the structure of the datasets used in the experiments. Section V presents the results of our experiments, and Section VI details the discussion and comparison of the methods used. Section VII presents the limitations, and finally, Section VIII concludes the paper and highlights potential avenues for future work.

II. RELATED WORK

The advent of predictive policing, which is part of smart policing, has been helpful in addressing crime-related challenges. Predictive policing relies on various techniques, including mapping techniques, ML, and NLP [14]. These techniques are employed to classify crime incidents based on their characteristics, predict the likelihood of crimes in specific locations, or identify individuals who are most likely to engage in criminal activities [6].

Mapping techniques form the foundation of crime prediction by enabling the identification of crime hotspots. These techniques include point mapping, thematic mapping of geographic areas, spatial ellipses, grid thematic mapping, and kernel density estimation (KDE) [15]. KDE is widely used for visualizing crime data because it effectively identifies hotspots without being constrained by geometric shapes such as ellipses [16].

ML has become a valuable tool for crime prediction tasks, such as identifying crime hotspots and predicting crime categories. Various ML techniques, including support vector machines, naïve Bayes, artificial neural networks, K-nearest neighbors, decision trees, and random forests, have been employed in this context [17]. A study provides a comprehensive analysis of ML methods for crime prediction using crime datasets from Chicago and Los Angeles. This study suggests that XGBoost is the best choice for crime prediction with an accuracy score of 94% for the Chicago dataset and 88% for the Los Angeles dataset [18]. In another study, naïve Bayes and back propagation neural networks were used to identify and predict crime categories based on socioeconomic datasets related to various locations in the United States. Their experiments showed that naïve Bayes outperformed neural networks with an accuracy of 94.08% [19], whereas, in a similar study on crime category prediction in the United States, decision trees outperformed naïve Bayes [20]. Decision tree is a popular and powerful ML algorithm in this context, as shown in [21] and [22], who compared decision tree j48 with a support vector machine, naïve Bayes, and neural network. The authors in [23] also suggested crime category prediction in specific geographical areas using naïve Bayes and decision tree based on historical incident data sourced from the Chicago Police Department's CLEAR system, both of which were applied to the top nine selected features from the dataset, and as a result, the decision tree outperformed naïve Bayes.

Clustering methods, such as K-means, have also been useful for identifying regions with high crime rates and predicting future criminal activities [24]. Additionally, regression methods, including Negative Binomial and Poisson Regression, have shown promise in crime prediction based on time-series data [25].

Deep learning methods, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks, have gained prominence in smart policing. These deep learning methods are used for criminal image classification, handling pooled cross-sectional data, and enhancing security [26], [27]. For instance, CNNs extract essential features from maps, whereas RNNs are useful for analyzing temporal structures within the data [28]. In a study that combined RNN and CNN to predict crime categories, the feedforward neural network achieved an accuracy of 71.3%, CNN achieved an accuracy of 72.7%, RNN achieved an accuracy of 74.1%, and their proposed method improved the accuracy by 1.5% [29].

Studies have also explored unsupervised learning and crime linkage using NLP. Crime linkage aims to identify crimes committed by the same individuals, whereas unsupervised NLP techniques have been employed to cluster crimes and inform policing strategies [30]. Moreover, there is emerging interest in utilizing NLP to analyze social media data, such as Twitter, to predict crime rates in cities [31].

In recent years, LLMs, such as GPT series models, have attracted interest and have drawn attention from law enforcement agencies for their potential use in smart policing. According to previous studies, while LLMs hold promise for supporting policing through NLP, ethical concerns must be addressed [14]. The selection of BART, GPT-3, and GPT-4 for our study was motivated by their distinct architectures and capabilities, which contributed to the comprehensive exploration of LLMs. GPT models offer a powerful contextual understanding, making them suitable candidates for tasks requiring the comprehension of textual information. GPT-3, or the "Generative Pre-trained Transformer 3," is a cutting-edge model belonging to the transformer architecture family. With an impressive scale of 175 billion parameters, it utilizes the same model structure as GPT-2, featuring alternating dense and sparse attention layers. Its extensive pre-training on diverse Internet text data enables it to effectively generalize across various domains, making it proficient in tasks ranging from article writing to question answering and code creation [11]. GPT-4 was introduced in March 2023 and significantly advanced by incorporating multimodal signal processing with text input. GPT-4 prioritizes safety in development practices, addressing concerns such as hallucinations, privacy, and overreliance with the introduction of intervention strategies such as red teaming. The model is built on a deep learning architecture based on the concept of predictable scaling, ensuring accurate performance forecasting with minimal computation during training and improved overall efficiency and predictability [32]. By including GPT-4 in our study, we aimed to assess the impact of model advancements on crime prediction performance in comparison with the GPT-3 model.

BART, short for bidirectional and auto-regressive transformers, is another LLM with a denoising autoencoder designed for pre-training sequence-to-sequence models [33]. BART employs a standard transformer-based neural machine translation architecture that is powerful in capturing contextual dependencies in textual data. Despite its simplicity, BART's design draws inspiration from and extends the capabilities of other prominent pre-training schemes, such as BERT [12] and GPT.

The capabilities of LLMs have been investigated across a range of domains, including agriculture [34], legal predictions [35], [36], [37], medical [38], [39] and financial applications [40], [41], utilizing various techniques for fine-tuning and prompting LLMs such as GPT-3 and BERT. However, the integration of LLM into smart policing remains unexplored. In this paper, we provide guidance for the design and development of an LLM-assisted framework in smart policing, and as a prototype, we perform an experimental analysis of using LLMs to classify crime occurrences based on crime categories and information from crime incident reports such as date, time, and location. Additionally, we compared the performance of the LLMs with that of the random forest ML algorithm as the baseline model for crime prediction.

III. METHODOLOGY

This section describes the architecture and operational mechanism of a novel framework designed to enhance smart policing through the application of LLMs to crime prediction and classification tasks. As depicted in Figure 1, this framework, systematically integrates LLMs into the predictive analytics domain, focusing on the critical steps required to adapt these models for smart policing applications, including but not limited to crime prediction.

The first step is the meticulous selection and preparation of a data source according to the defined task (e.g. crime prediction), thereby ensuring that the foundational data are both relevant and optimized for processing. Data sources can be datasets of information extracted from police reports in local law enforcement agencies in different areas or cities. After applying pre-processing techniques, including addressing missing values through imputation or dropping, prompt engineering is used to transform the original historical crime dataset into a human-language-understandable format; therefore, each data point is a description of a crime incident in natural language. Subsequent steps involve the creation of an instruction dataset derived from the original dataset, to serve as the basis for fine-tuning LLMs. Additionally, the interaction with LLMs is conducted via prompt engineering through their Application Programming Interface (API), which is a critical process that enables the customization of model responses to align with the specific needs of a crime prediction system.

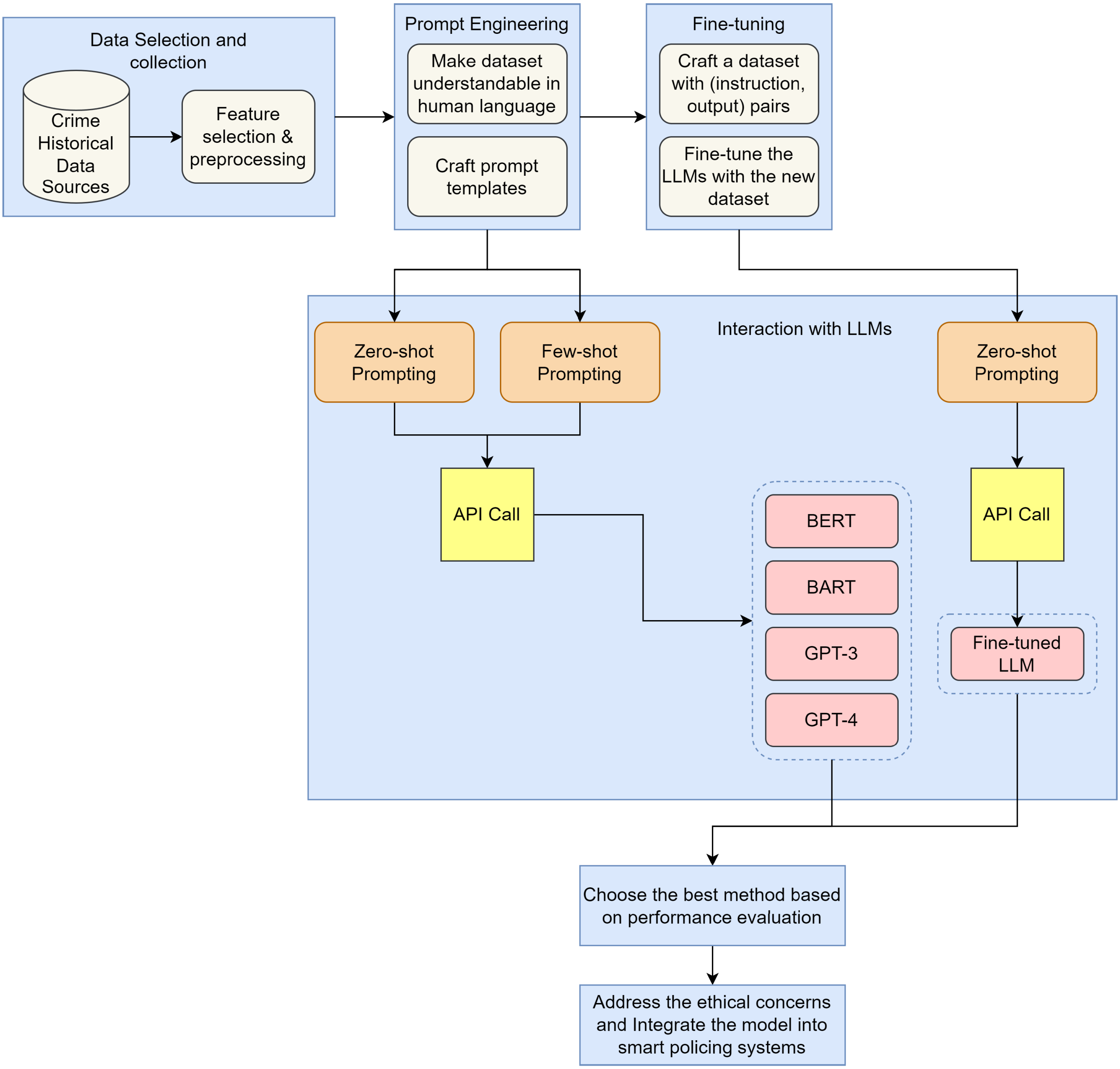

Data Selection and collection
Crime Historical Data Sources
Feature selection & preprocessing
Prompt Engineering
Make dataset understandable in human language
Craft prompt templates
Fine-tuning
Craft a dataset with (instruction, output) pairs
Fine-tune the LLMs with the new dataset
Interaction with LLMs
Zero-shot Prompting
Few-shot Prompting
Zero-shot Prompting
API Call
BERT
BART
GPT-3
GPT-4
API Call
Fine-tuned LLM
Choose the best method based on performance evaluation
Address the ethical concerns and Integrate the model into smart policing systems

FIGURE 1. **Framework architecture for developing LLM-assisted tools for smart policing.**

Upon gathering the results from interactions with LLMs, a comprehensive performance evaluation was performed. This evaluation process is crucial for identifying the most effective method for crime classification and prediction and integrating the crime prediction tool into smart policing systems.

In this paper, a spatiotemporal crime prediction tool based on LLMs was designed and tested with two sets of historical crime data as prototypes. As shown in Figure 2, the tool, indicated in yellow, can operate within a broader smart policing system with diverse data types for inputs from different sources to generate more accurate predictions and actionable insights into criminal patterns and predictive analytics for law enforcement use. However, the investigation of other modules and their integration into a single system is left for future work.

Users such as police officers can interact with smart policing systems by entering the required input for each module. Using data visualization tools, including dashboards and heat maps alongside the crime prediction module, they can understand what type of crime activities are more likely to happen in specific locations and times. Therefore, they can decide how to allocate security resources or prioritize their patrolling in different locations. The system architecture allows for flexible deployment on either cloud-based services or local devices, depending on the specific requirements of

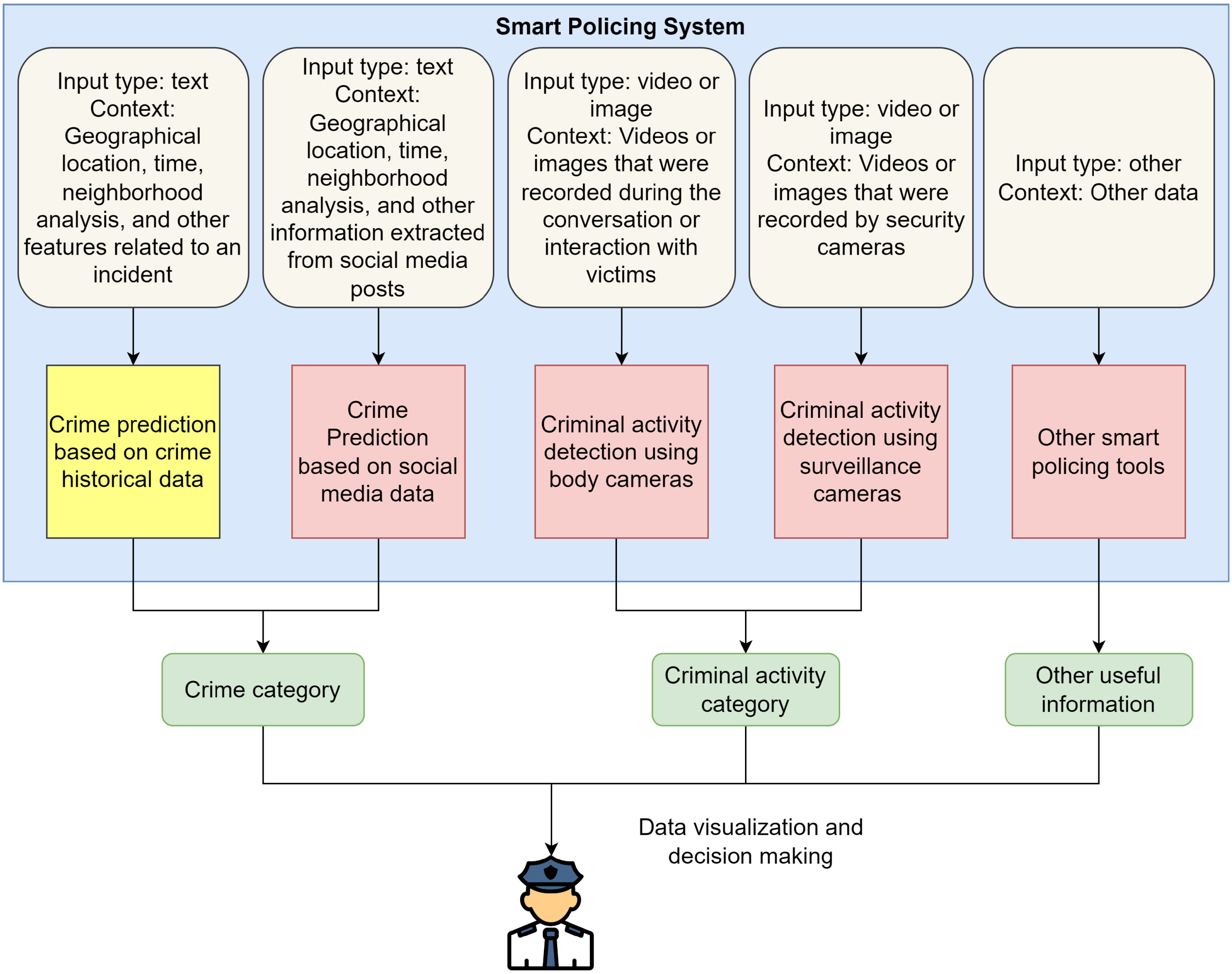

Smart Policing System
Input type: text Context: Geographical location, time, neighborhood analysis, and other features related to an incident
Input type: text Context: Geographical location, time, neighborhood analysis, and other information extracted from social media posts
Input type: video or image Context: Videos or images that were recorded during the conversation or interaction with victims
Input type: video or image Context: Videos or images that were recorded by security cameras
Input type: other Context: Other data
Crime prediction based on crime historical data
Crime Prediction based on social media data
Criminal activity detection using body cameras
Criminal activity detection using surveillance cameras
Other smart policing tools
Crime category
Criminal activity category
Other useful information
Data visualization and decision making

FIGURE 2. Integration of an LLM-Assisted framework (indicated by yellow) into a smart policing system and how users can interact with them.

the policing context and available computational resources. Cloud deployment, for instance, enables real-time updates with new data, thereby enhancing the system's effectiveness in identifying new crime trends.

A. PROMPT ENGINEERING AND FINE-TUNING

In this paper, prompting and fine-tuning methods were used to interact with LLMs, such as GPT models and BART, to analyze their abilities in crime classification and prediction tasks. Practically, this was done through the OpenAI API for GPT models and the HuggingFace repository pipeline function for BART with their default parameters settings, and fine-tuning through instruction tuning was used to adopt GPT models into the domain. Our choice of GPT models and BART is driven by their power in downstream tasks and specific architectures. Additionally, as previously mentioned, these models have been extensively studied and considered for adoption across diverse applications, highlighting their versatility and potential efficiency.

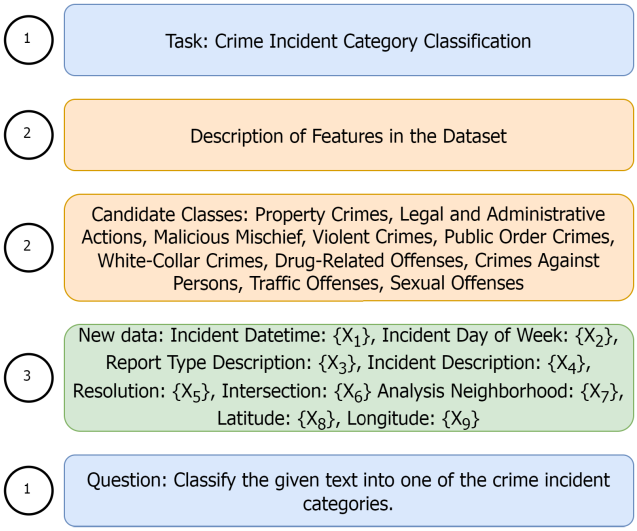
The effectiveness of prompting LLMs is tied to the quality and comprehensiveness of prompts as they guide the model's predictions. Although simple prompts can be efficient for straightforward cases, providing additional information about the task and desired output format can be beneficial for achieving optimal performance. The use of prompts is often associated with the manual engineering of prompts, in which we must carefully design and evaluate each element of the prompt to achieve the desired results in the crime prediction task. Our investigation of the application of zero-shot and few-shot prompting techniques to the BART and GPT models allowed us to explore their adaptability in predicting crime categories beyond their explicit training data without limiting their prior knowledge to a specific domain.

In the realm of ML, zero-shot learning implies that a model is trained to recognize and classify instances that have never been seen before. This is typically achieved by using semantic relationships and attributes to generalize knowledge from seen or known classes to unseen or unknown classes. Unlike traditional ML models, in which systems are trained on a dataset and subsequently tested on unfamiliar data within the same class, zero-shot learning empowers models to predict unobserved classes, thereby enabling systems to generalize knowledge and make predictions beyond their initial training data [42], [43].

In the context of LLMs, zero-shot learning takes on a specific form in which text-based descriptions in prompts are used to enable the model to perform tasks; it is not explicitly trained. This process involves employing text prompts that guide the model in understanding its tasks, dataset, and expected inputs or outputs, and predicting classes of crime incidents, leveraging pre-existing knowledge of LLMs for new fields and applications. Additionally, this approach enables the model to make predictions, even in scenarios where there is a lack of annotated data [44].

For prompting in a zero-shot manner, we considered different templates to make the data understandable for LLM. First, a prompt template with incomplete text or masked slots is applied. Subsequently, the knowledge acquired by the pre-trained language model is used to predict the token that can fill in the slot, complete the text, or answer a question. To mitigate prompt engineering challenges, we used a class of simple prompts known as null prompts. As shown in previous studies, these null prompts, which involve simple concatenations of inputs and the masked token, demonstrate comparable accuracy to manually written patterns while significantly simplifying the prompt design [45], [46]. The design of such prompts is similar to that of question-answer prompts. We assessed the outputs generated by the LLMs by focusing on their relevance, accuracy, and complementarity with our objectives. This iterative evaluation process was continued, improving the prompt, until the LLMs consistently produced results that aligned closely with our predefined standards and expectations of accuracy and contextual relevance. Finally, we used the prompt template shown in Figure 3. Part (1) of the template specifies the task description of the model. Part (2) contains information regarding the dataset, such as its features and label classes. Finally, part (3) provides the features of the new data that the model should classify.

Few-shot learning in LLMs involves an approach similar to zero-shot learning that uses prior knowledge of the LLM. However, in a few-shot manner, the model is provided with a small number of examples for the defined task without updating its internal weights. Each example typically consists of the input data and the desired outcome, such as an English sentence and its corresponding French translation. To perform few-shot prompting, K examples of input data and outcomes are provided, followed by a new input, and the model is expected to generate an appropriate outcome [11]. We used the same template as in Figure 3 for few-shot prompting LLMs, in addition to two examples of each class. In this case, examples include the characteristics of a crime incident and the corresponding crime category of that incident.



Task: Crime Incident Category Classification

Description of Features in the Dataset

Candidate Classes: Property Crimes, Legal and Administrative Actions, Malicious Mischief, Violent Crimes, Public Order Crimes, White-Collar Crimes, Drug-Related Offenses, Crimes Against Persons, Traffic Offenses, Sexual Offenses

New data: Incident Datetime: $\{X_1\}$, Incident Day of Week: $\{X_2\}$, Report Type Description: $\{X_3\}$, Incident Description: $\{X_4\}$, Resolution: $\{X_5\}$, Intersection: $\{X_6\}$ Analysis Neighborhood: $\{X_7\}$, Latitude: $\{X_8\}$, Longitude: $\{X_9\}$

Question: Classify the given text into one of the crime incident categories.

FIGURE 3. **Prompt template that we used for zero-shot prompting and few-shot prompting.**

Fine-tuning finds applications in domain adaptation, enhancing a model's performance on target data with different distributions, and data augmentation, where transformation of existing data creates new samples to improve model performance, especially with limited labeled data. Fine-tuning language models as depicted in Figure 4, refers to the adaptation of a new specific task by further training a pre-trained model on that task or its relevant dataset. Istruction fine-tuning, also known as supervised fine-tuning involves updating the model's weights and parameters based on a small task-specific dataset while leveraging the general language understanding and generation capabilities learned during the pre-training phase [47]. This method is identified as a process for acquiring the instructions-following capability of LLMs and enhancing the controllability of LLMs, as recognized in various real-world scenarios rather than the typical next-word prediction objective of language models [48], [49].

The construction of a domain-specific dataset or instruction dataset involves two main steps of data integration, in which (instruction, output) pairs are collected from existing annotated datasets by transforming text-label pairs into (instruction, output) pairs and generating outputs where outputs are obtained by employing LLMs, such as GPT-3.5-Turbo or GPT-4, based on manually collected or expanded instructions.

The pre-trained model's parameters serve as the starting point of the fine-tuning. These include weights and biases that the model has adjusted during its initial training phase on the large dataset. The purpose of fine-tuning is to adjust these parameters slightly to suit the new task without losing the generalized understanding the model has acquired. Then, the model is then trained or fine-tuned on the task-specific dataset. This training process does not start from scratch but rather adjusts the pre-trained parameters. The learning rate used during fine-tuning is typically much lower than that

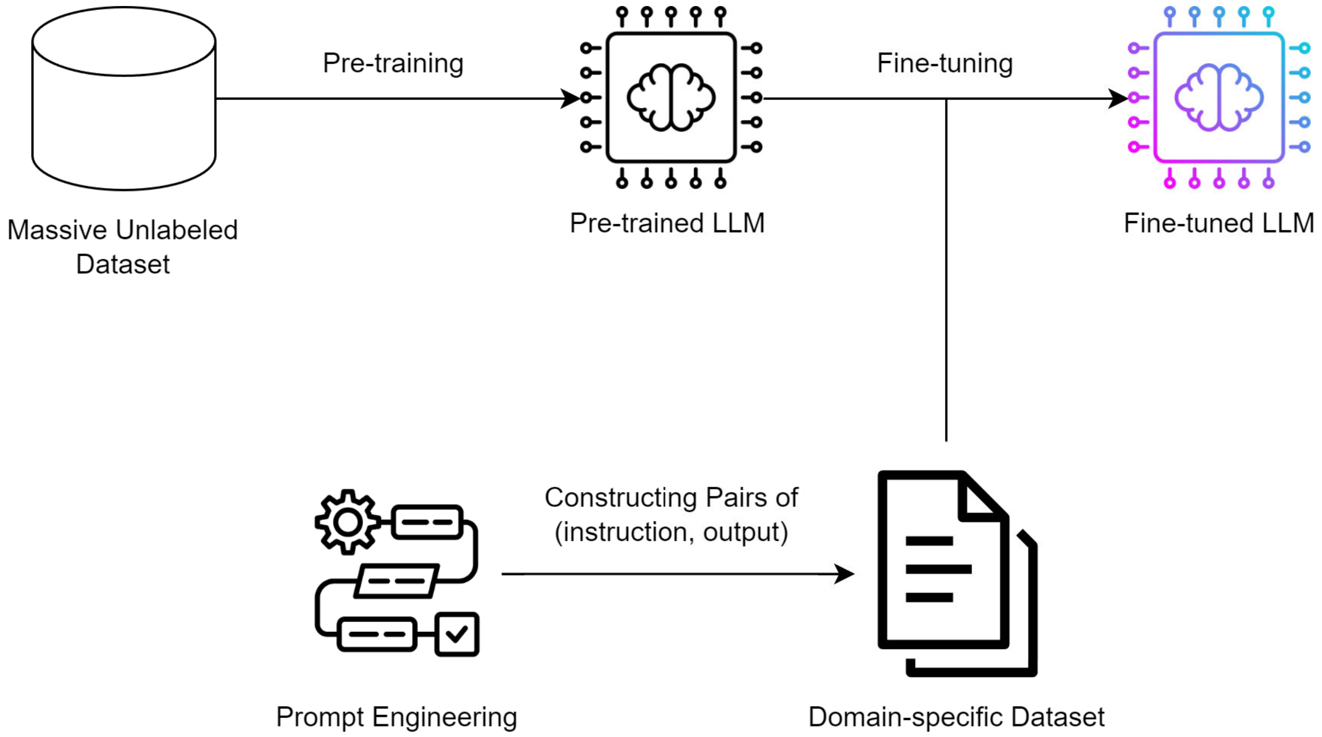


FIGURE 4. The Fine-tuning process through instruction tuning of LLMs.

used in the initial training phase to ensure that the pre-learned patterns are not drastically altered. The power of instruction fine-tuning lies in its ability to refine the model's responses, ensuring they are not only accurate but also contextually relevant and aligned with the defined task of crime prediction enabling these advanced models to interpret and respond to complex crime data.

In this paper, we fine-tuned the GPT-3.5 using the OpenAI platform's instruction fine-tuning tool which uses computational frameworks to manage the complexities of backpropagation, gradient descent, and parameter updates. For this purpose, we created an instruction dataset based on the original dataset of crime incidents in Los Angeles and San Francisco to classify crime incidents based on incident categories.

B. ML MODELS

To compare the ability of the LLMs and ML models in classification and prediction tasks, we considered the random forest (RF) and XGBoost models as the baselines of our experiments. RF is an ensemble learning model comprising multiple decision trees, each constructed independently with a unique set of randomly selected features. The ensemble model aggregates the results of these individual tree classifiers through a voting mechanism to make the final decision. The randomness in RF has two key aspects: random selection of training samples using the bagging algorithm and random choice of split attributes in each internal node [50]. RF has the advantage of handling imbalanced datasets where the number of instances in different classes is uneven. The ability to differentiate the cost ratio between false negatives and false positives allows for a more meaningful evaluation of predictive accuracy. This approach ensures that the model considers the real-world consequences of false negatives and false positives, and addresses the imbalance of our datasets in the outcome variable, which increases its practical utility.

XGBoost, also known as Extreme Gradient Boosting, is another ensemble learning model that represents an advanced ML algorithm within the gradient boosting family specifically designed for exceptional model performance and computational efficiency. XGBoost sequentially trains decision trees on the training data, and iteratively adds new trees to adjust the errors made by existing ones. The algorithm optimizes an objective function that combines a loss term, measuring prediction errors, and a regularization term to control model complexity and prevent overfitting. XGBoost uses a second-order Taylor expansion to approximate the loss function and determines the optimal tree structure using a greedy algorithm. Notable advantages of XGBoost include internal handling of missing data, elimination of the need for imputation, high computational speed through parallel processing, and robustness against overfitting. The versatility and ability of the algorithm to deliver high prediction accuracy make it a preferred choice among researchers and practitioners in the field of ML [51], [52].

IV. DATA COLLECTION

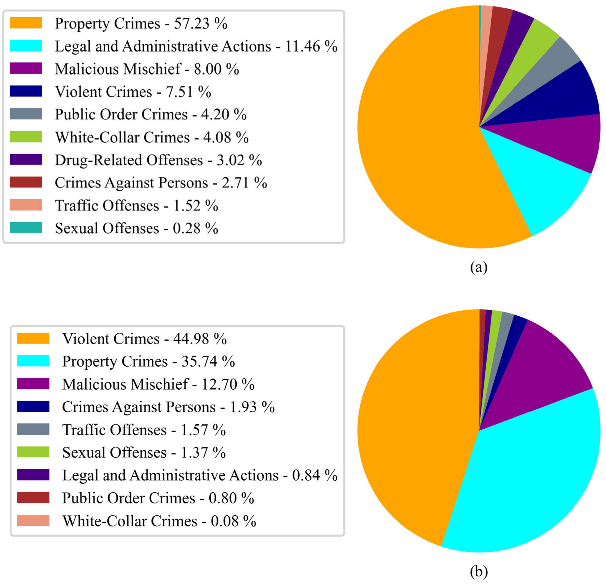
To perform our experiments, we used two popular datasets in the context of crime prediction: crime data from San Francisco (SF) and Los Angeles (LA). To facilitate our

TABLE 1. Datasets' features and description.

Dataset	No.	Feature Name/API	Description	No. of Null Values
SF	1	Incident Datetime	The date and time when the incident occurred	0
	2	Incident Day of Week	The day of the week the incident occurred	0
	3	Report Type Description	The description of the report type	0
	4	Incident Category	A category mapped onto the Incident Code	176126
	5	Incident Description	The description of the incident	0
	6	Resolution	The resolution of the incident	0
	7	Intersection	The 2 or more street names that intersect closest	42357
	8	Analysis Neighborhood	The neighborhood where each incident occurs	42506
	9	Latitude	The latitude coordinate in WGS84	42357
	10	Longitude	The longitude coordinate in WGS84	42357
LA	1	DATE OCC	The date that the crime occurred (MM/DD/YYYY format)	0
	2	TIME OCC	The time that the crime occurred (24-hour military time)	0
	3	AREA NAME	The geographic area or patrol division name	0
	4	Crm Cd Desc	Description of the crime committed	0
	5	Vict Age	Two-character numeric representing victim age	0
	6	Vict Sex	F - Female, M - Male, X - Unknown	113250
	7	Vict Descent	Descent Code: A - Other Asian, B - Black, C - Chinese, ...	113258
	8	Premis Desc	Type of structure, vehicle, or location of the crime	523
	9	Weapon Desc	Type of weapon used in the crime	559467
	10	LOCATION	Street address of the crime incident (rounded to the nearest hundred block)	0
	11	LAT	Latitude of the crime incident	0
	12	LON	Longitude of the crime incident	0

TABLE 2. Crime incident categories.

Incident category after pre-processing	Original Incident categories
Property Crimes	Larceny Theft, Burglary, Motor Vehicle Theft, Recovered Vehicle, Stolen Property, Robbery, Lost Property
Violent Crimes	Assault, Homicide, Rape
White-Collar Crimes	Fraud, Embezzlement, Forgery and Counterfeiting
Public Order Crimes	Disorderly Conduct, Arson, Vandalism, Liquor Laws, Gambling, Weapons Offense, Weapons Carrying
Crimes Against Persons	Crimes Against The Family And Children, Suicide, Missing Person
Drug-related Offenses	Drug Offenses, Drug Violation
Sexual Offenses	Sex Offense, Prostitution, Human Trafficking, Commercial Sex Acts
Traffic Offenses	Traffic Violation Arrest, Traffic Collision, Vehicle Impounded
Legal and Administrative Actions	Non-Criminal, Warrant, Case Closure, Courtesy Report
Malicious Mischief	Malicious Mischief



(a)

(b)

FIGURE 5. Data distribution; (a) Data distribution for the SF dataset, (b) Data distribution for the LA dataset.

analysis, we meticulously pre-processed the data to address the missing values. The features used in our experiments are listed in Table 1. In this table, the features, along with their descriptions and corresponding number of null values, provided essential insights into the structure of the dataset. For instance, the "Vict Sex" feature in the LA dataset indicates the gender of the victim, while "Weapon Desc" details the type of weapon used in the crime. We observed a substantial number of null values in the "Vict Descent" feature, indicating unknown or unreported ethnicities of victims. We attempted to handle null values by dropping data or replacing them with sufficient descriptions. For example, for the "Weapon Desc" feature, we assumed that no weapons were used in crime occurrence.

The first dataset included more than 750,000 reports of crime incidents in SF City from 2018 to the present collected from [53]. This dataset has 27 features and 41 unique incident types that were used for the incident classification. We selected ten features and mapped the incident categories into a more manageable set of classes during the pre-processing phase. The transformation is detailed in Table 2, and the graphical representation in Figure 5 (a) illustrates the uneven distribution of data across incident type categories with Property Crimes being the most prevalent crime, indicating the dataset's imbalance distribution. Understanding data distribution is crucial for comprehending a dataset's characteristics, imbalances, and potential challenges in crime classification.

The second dataset contains information on crime reports in LA dating back to 2020 [54]. This dataset includes more than 850,000 records of crime occurrences in LA City with 28 features. We selected 12 features, as shown in Table 1, and set crime type as the label. Similar to the SF dataset, we transformed the crime types into a broader set of categories, as specified in Table 2. Figure 5 (b) also visually represents the data distribution of the LA dataset. The distribution shows the prevalence of different crime categories, with Property Crimes and Violent Crimes being the most prominent. This comprehensive exploration of the features and distribution lays the foundation for subsequent analysis and model evaluation.

V. EXPERIMENTAL RESULTS

In this section, we provide a comprehensive experimental analysis that applies both LLMs, including the BART and GPT models, and traditional ML models, namely RF and XGBoost, to two crime datasets from San Francisco (SF) and Los Angeles (LA). Before the analysis, these datasets were pre-processed to align with the requirements of each model, focusing on crime prediction and classification tasks. To evaluate the performance of each model, the datasets were partitioned into training and test sets, allocating 80% of the data for training and the remaining 20% for testing. It is important to note that the same subsets of data were utilized to train the ML models and fine-tune the GPT-3 model. Given the constraints related to the size and computational costs associated with fine-tuning GPT-3 across the entire training dataset, a reduction in the size of the training set was considered.

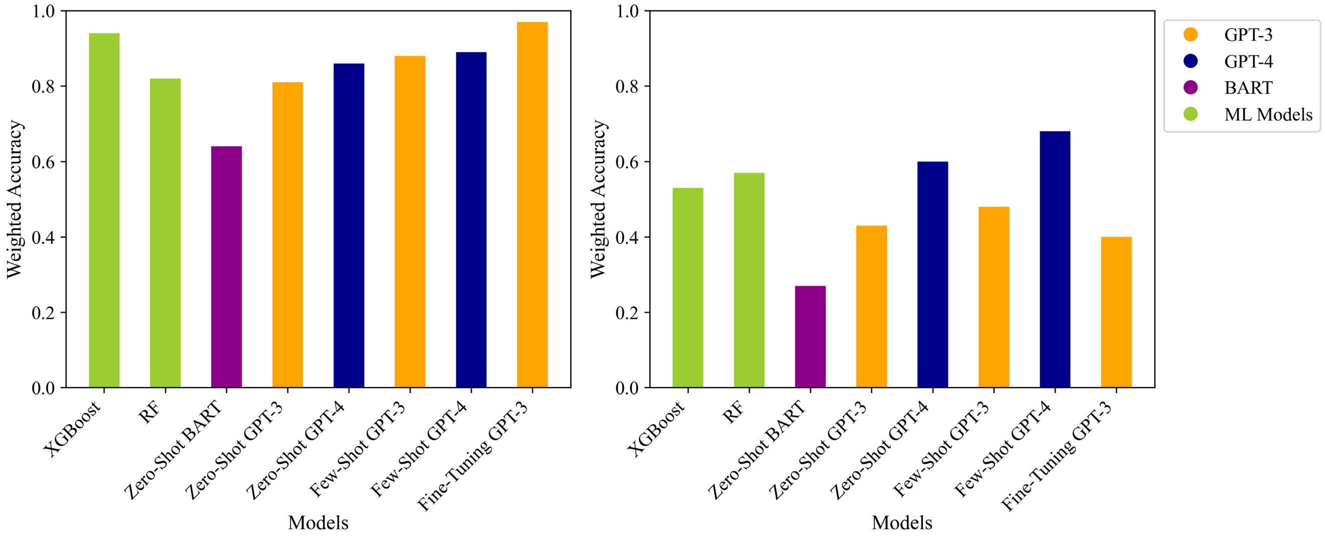
Due to the imbalanced nature of the datasets, the evaluation metrics included the weighted averages of accuracy, precision, recall, and F1-score. These metrics were calculated to assess the model performance, effectively accounting for disparities in the class distribution. In this context, precision measures the proportion of correctly predicted positive observations to the total predicted positives for a specific crime category. High precision for a specific crime type suggests that the model can effectively identify crime types without confusing them with others. On the other hand, recall measures the proportion of actual positives correctly identified by the model for a specific crime category. High recall indicates that the model can identify most actual instances of a specific crime type. Class imbalance is a common issue in crime datasets in which some crime types are more frequent than others. Therefore, precision and recall were used to highlight how well a model performed for both majority or frequent crime and minority or rare crime categories, guiding efforts to improve model sensitivity and specificity across the board. Additionally, the F1-score considers the balance between precision and recall and is particularly useful in scenarios where both false positives and false negatives carry significant consequences, as is often seen in crime prediction and classification tasks. The results of these evaluations are graphically represented in Figures 6 and 7, and a detailed breakdown of the performance metrics for both the minority and majority classes is presented in Table 3.

A. SAN FRANCISCO DATASET

The results obtained from the experiments for the crime prediction task on the SF dataset reveal significant insights into the performance of each approach. As shown in Figures 6(a) and 7(a), the baseline ML models, RF and XGBoost, exhibited good performance, achieving weighted accuracies of 82% and 94%, respectively. Comparatively, the Fine-tuned GPT-3 model surpassed these benchmarks with an accuracy score and F1-score of 97%. Although XGBoost as a supervised ML algorithm demonstrated robust performance, it did not reach the benchmark set by the fine-tuned GPT-3 model, highlighting the unique strengths of LLMs that have undergone fine-tuning for specific classification tasks. The good performance of GPT-3 can be attributed to its ability to leverage vast amounts of pre-trained knowledge and adapt this understanding to crime classification and prediction within the SF dataset. This suggests that for tasks involving complex pattern recognition and classification within substantial and diverse data sets, fine-tuned language models offer a more effective approach than traditional ML models.

An interesting observation is that zero-shot prompting on GPT-3 demonstrated performance closely aligned with that of the RF model. In contrast, BART showed less favorable results, whereas zero-shot prompting on GPT-4 surpassed both. This distinction underscores the inherent strengths of GPT models, which, as auto-regressive language models pre-trained on vast corpora, exhibit a significantly better contextual understanding relative to BART's bidirectional architecture, which primarily focuses on capturing contextual dependencies. As shown in Figure 6, zero-shot prompting with BART results in a high precision rate which signifies that when BART classifies an incident under a certain crime category, it is highly likely to be correct. This is evidence of the model's effectiveness in correctly interpreting the features that define each crime category, minimizing false positives when incidents are wrongly labeled under a category. Conversely, the low recall of zero-shot prompting BART highlights a limitation in the model's capability to identify all actual instances of a given crime category, where BART misses a significant number of real occurrences of certain crimes, failing to label them as such.

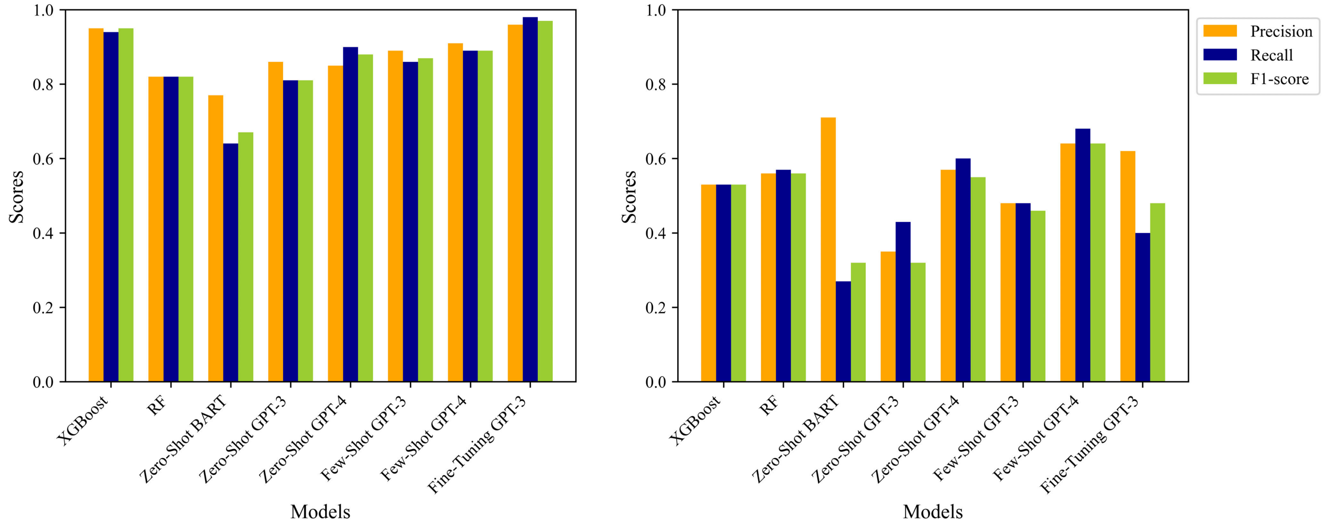
Moreover, few-shot prompting improved the performance of both the GPT-3 and GPT-4 models achieving notable improvements in precision, recall, and F1-score. These results demonstrate the critical role of incorporating specific examples of data during training with an equal number of examples for each class, thereby improving the model's performance and accurately classifying frequent and rare instances. Additionally, the observed slight increase in the



(a)

(b)

FIGURE 6. Weighted accuracy for different methods; (a) Results for the SF dataset, (b) Results for the LA dataset.



(a)

(b)

FIGURE 7. Comparison of methods based on precision, recall, and F1-score; (a) Results for the SF dataset, (b) Results for the LA dataset.

weighted average of accuracy and the F1-score suggests that few-shot prompting aids in the model's overall generalization capabilities.

According to the evaluation results presented in Table 3, the zero-shot prompting method employed by the GPT-3 model showed robust performance on the majority class for the SF dataset, as evidenced by the high precision, recall, and F1-score metrics. However, this model showed a notable deficiency in identifying instances of the minority class, as indicated by the zero predictive accuracy for this group. This disparity highlights the limitations of the model in generalizing to minority classes in the absence of explicit training examples, highlighting a critical challenge in deploying zero-shot prompting methods for balanced classification tasks. In contrast, zero-shot prompting on GPT-4 significantly outperformed GPT-3 and other methods in both minority and majority class predictions for the SF dataset, achieving perfect scores. This illustrates GPT-4's advanced ability to leverage its extensive pre-training for a more effective generalization across diverse class distributions, even without providing examples of the training dataset and fine-tuning.

In both zero-shot and few-shot scenarios, GPT-4 showed a marginal accuracy enhancement over GPT-3. Despite achieving perfect scores in precision, recall, and F1-score for the minority class, GPT-4 displayed a minor decline in performance metrics for the majority class. This pattern suggests that, while GPT-4 can identify minority class instances with high accuracy, it may not generalize across class distributions as effectively as GPT-3, particularly for the majority class. Both GPT-3 and GPT-4 showed high performance in few-shot scenarios, demonstrating a relatively high precision, recall, and F1-score for both minority and majority classes. It is interesting that while GPT-4 performs better in classifying minority and majority classes than GPT-3, its performance in identifying these classes slightly decreased in a few-shot manner according to its precision. This result suggests that enhancements in GPT-4 may not substantially affect the overall model generalization for the SF dataset. In general, although GPT-4 offers some improvements through zero-shot and few-shot prompting, the fine-tuned GPT-3 model remains the most powerful model for crime classification on the SF dataset.

B. LOS ANGELES DATASET

The analysis of the LA dataset revealed distinct challenges in crime prediction, presenting notable performance disparities when compared with the SF dataset. According to the results shown in Figure 6 (b) and Figure 7 (b), the XGBoost and RF models as baselines exhibit weighted accuracies of 53% and 57% with F1-score of 53% and 56% respectively. This represents a considerable decline in their effectiveness relative to their performance in the SF context. Moreover, the fine-tuned GPT-3 model exhibited a weighted accuracy of 40%, indicating a pronounced decrease compared to its counterpart within the SF dataset. Such disparities underscore the dataset-specific performance of both ML models and LLMs, and the necessity for analytical approaches that meticulously consider the unique attributes of crime incidents inherent to each geographical location.

Zero-shot prompting on GPT-3 and GPT-4 demonstrated varying performance on the LA dataset. GPT-3 showed limited effectiveness, whereas GPT-4 achieved a comparatively higher accuracy of 60%. This variance in performance can be attributed to GPT-4's more advanced architecture and larger training corpus, which likely provides it with better contextual understanding and adaptability to new datasets. Consistent with earlier observations, BART displayed less favorable outcomes with significantly high precision but low recall, possibly because of its bidirectional architecture's limitations in adapting to the specific linguistic and structural features of crime data without direct examples.

Similar to the results for the SF dataset, few-shot prompting contributed to the performance enhancement for both GPT models on the LA dataset as it achieved an accuracy of 43% for GPT-3 and 68% for GPT-4. However, it still lags behind. This suggests that few-shot prompting is still not sufficient to address the complexities and imbalances of crime occurrences within the LA dataset, but there are still improvements compared to the RF and XGBoost models as baselines. This suggests that the LA dataset poses unique challenges, potentially related to its inherent characteristics such as class distribution, complexity of crime descriptions, or other contextual factors not as prevalent in the SF dataset. Despite a few explicit examples through few-shot prompting, the complexity and possibly greater heterogeneity of the LA dataset may limit the effectiveness of this approach.

Analyzing the performance of methods in distinguishing majority and minority classes according to the results in Table 3, zero-shot prompting on GPT-3 and GPT-4 exhibited limited effectiveness in predicting minority class instances, with both models failing to identify any instances correctly, as indicated by their zero precision, recall, and F1-score. In contrast, zero-shot GPT-3 slightly outperformed GPT-4 in majority class predictions, achieving a higher F1-score of 64% compared to GPT-4's 60%, which contradicts earlier expectations based on their performance on the SF dataset. This outcome underscores the challenges posed by the LA dataset, which may include more complex or varied crime descriptions that are difficult to generalize from pre-trained knowledge without explicit examples.

Few-shot prompting offered modest performance improvements in identifying minority and majority class predictions for both GPT models, with GPT-4 showing a slight edge over GPT-3 in terms of F1-score. However, the enhancements in minority class prediction were not as significant as those observed in the SF dataset. It is noteworthy that BART and traditional ML models such as RF and XGBoost also faced challenges, with RF and XGBoost unable to identify minority class instances accurately, as shown by their zero values in precision, recall, and F1-score for the minority class. These results highlight the dataset-dependent nature of model performance, underscoring the complexities of crime prediction tasks and the necessity for adaptable, context-aware modeling approaches to tackle the specific characteristics of crime incidents within diverse urban settings.

Surprisingly, the fine-tuned GPT-3 model experienced a significant drop in performance compared to its results on the SF dataset, achieving a weighted accuracy of 40% and F1-score of 48% with a substantial decrease in recall as shown in Figure 6 (b) and Figure 7 (b). Despite experiencing a relative decrease in accuracy, GPT-4 has emerged as a noteworthy option for crime prediction within the LA dataset. However, the marginal improvements introduced by GPT-4 do not unequivocally establish it as a superior choice for this task.

The challenges faced by GPT-4 and fine-tuned GPT-3 in generalizing to the minority class within the LA dataset highlight the need for enhanced strategies capable of navigating the diverse and complex nature of urban environments. This suggests the crucial role of model selection, emphasizing that advancements in model architecture require careful consideration to effectively leverage their full potential. The observed performance variations between GPT-3 and GPT-4 in the LA dataset suggest that a deeper understanding of each model's strengths and limitations is essential for optimizing crime prediction efforts across different geographical settings.

TABLE 3. Class prediction evaluation results for minority and majority classes.

Dataset	Methods	Minority Class Prediction			Majority Class Prediction		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
SF	Zero Shot BART	33	100	50	91	70	79
	Zero Shot GPT-3	0	0	0	93	92	93
	Zero Shot GPT-4	100	100	100	98	86	92
	Few Shot GPT-3	100	100	100	94	97	95
	Few Shot GPT-4	100	100	100	94	95	94
	Fine-Tuned GPT-3	100	96	98	100	100	100
	RF	38	23	29	73	62	67
	XGBoost	92	92	92	92	92	92
LA	Zero Shot BART	17	100	29	100	14	25
	Zero Shot GPT-3	0	0	0	49	91	64
	Zero Shot GPT-4	0	0	0	60	83	60
	Few Shot GPT-3	14	19	16	65	86	74
	Few Shot GPT-4	15	48	23	72	75	74
	Fine-Tuned GPT-3	14	31	19	69	65	68
	RF	0	0	0	68	55	61
	XGBoost	0	0	0	65	57	61

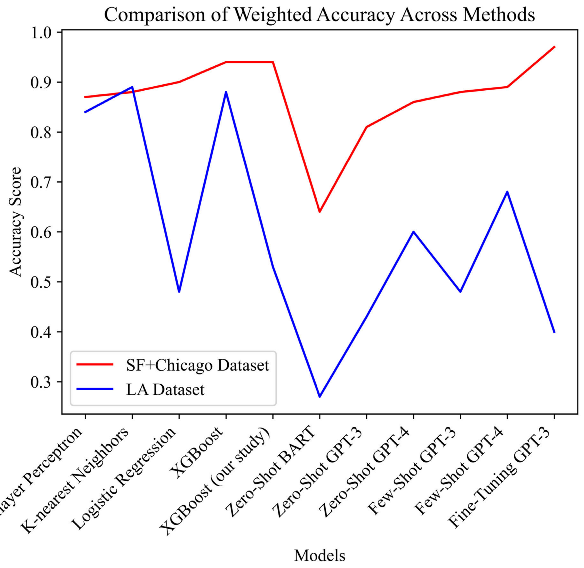


FIGURE 8. Comparison of oss ML methods including multilayer perceptron, K-nearest neighbors, logistic regression, and XGBoost in related work [18] and LLMs in our work.

VI. DISCUSSION

The experimental results show that LLMs can be used in smart policing specifically, crime prediction and classification. Although the performance of LLMs in crime classification tasks is better than traditional ML in most scenarios, some cases, particularly on the LA dataset, show controversial results, revealing that more experiments in different scenarios are required to build more practical tools and services. Additionally, by comparing our models with a broader spectrum of traditional ML methods, including multilayer perceptron, K-nearest neighbors, logistic regression, and XGBoost from related studies, we attempted to shed light on the relative strengths and limitations of various crime prediction approaches. The evaluation using the weighted accuracy scores in Figure 8 reveals performance variations across the different methods and datasets. To evaluate the results from the SF dataset, we compared it with the models' performance on the Chicago dataset, as the overall performance of the models reached high accuracy on this dataset. This comparison shows that our Fine-Tuned GPT-3 has an impressive weighted accuracy of 97%, surpassing all the other ML models or LLM on SF and LA datasets in addition to the ML models' performance on the Chicago dataset from the past. This is evidence of the potential of LLMs, particularly when fine-tuned for crime classification tasks. However, it is crucial to recognize that the choice of LLMs and the efficacy of fine-tuning depend on the specific characteristics of the dataset. The LA dataset poses unique challenges, as is evident from the lower overall performance scores. While traditional methods, such as K-nearest Neighbors, show higher accuracy in this context, LLMs demonstrate the feasibility of leveraging large-scale pre-trained models for more accurate crime prediction. It is also noteworthy that zero-shot learning, in which models make predictions for classes not observed during training, provides an interesting perspective on the adaptability of language models. In the SF dataset, the zero-shot GPT-3 and GPT-4 performed reasonably well, indicating the capacity of these models to generalize to new crime categories without further training. Few-shot prompting enhances the performance across all models, underscoring the importance of providing specific examples during training to improve the model's ability to classify diverse instances. In the LA dataset, few-shot prompting GPT 4 contributes to significant improvements in comparison to ML models but remains insufficient to overcome the complexities of crime occurrences in this specific urban environment.

One notable point of discussion arises from the unexpected decrease in the accuracy observed in our Fine-Tuned GPT-3 model on the LA dataset. Although related studies have reported a high accuracy of 89% with the K-nearest neighbor algorithm, our GPT-3 model achieved a lower accuracy of 40%. Some potential factors contributing to the surprising disparities in model performance could be the unique characteristics of crime incidents and distinct patterns in criminal activity in LA city, or an inherent bias within the dataset. Therefore, a deeper investigation of the underlying reasons is crucial for improving the generalization capabilities of LLMs across diverse urban environments. Understanding these local intricacies is vital for fine-tuning LLMs or ML models to ensure their effectiveness across different geographical contexts. Another interesting observation is the comparison of the performance of XGBoost in previous studies with the results of our work, as illustrated in Figure 7. This highlights the influence of various parameters on the performance of AI models and the importance of considering model implementation when conducting comparisons, particularly in terms of accuracy and F1-score.

Additionally, because of the performance results of the models in identifying frequent and rare crime incidents and the diversity of crime types and their unique characteristics, customizing LLMs for specific categories of crime could improve prediction accuracy. Developing models for particular crime incident types, such as property crimes, violent crimes, or malicious mischief, could allow for more accurate and effective predictions.

Furthermore, ethical concerns regarding integrating such methods into predictive policing and the potential biases should be addressed. The potential of these technologies to enhance public safety and operational efficiency is considerable; however, the consequences of their deployment warrant thorough ethical examination. A primary ethical concern lies in the risk of perpetuating biases present in the historical crime data. AI models, including LLMs, are only as unbiased as the data they are trained on. Given the documented disparities in policing and judicial processes, there is a real risk that these technologies could reinforce existing prejudices, leading to discriminatory practices against some communities. Additionally, collecting vast amounts of data, including personal information, and its sharing through interactions with an LLM raises concerns about individuals' rights to privacy. This concern can be addressed using LLMs that are accessible through personal computers or local edges instead of cloud-based LLMs such as GPT models. Police departments and law enforcement agencies should make this choice based on the regulations and the computational resources of their local devices.

Another ethical concern is accountability and transparency. Decisions made by LLMs, if not explainable, can undermine trust in law enforcement agencies. Any AI system used in policing must be accompanied by mechanisms that ensure decisions are transparent, interpretable, and, most importantly, subject to human oversight. This includes ongoing monitoring and transparency in model decision-making as visualized by police officers within police departments and law enforcement agencies. According to the special role of law enforcement agencies and how they interact with societies, an interdisciplinary strategy, engaging experts from policing, computer science, law, and ethics, is essential to tackle the challenges and operational needs associated with the deployment of algorithms, especially generative AI such as LLMs in smart policing. This approach defines guidelines for transparency, comprehensibility, and ethical considerations. These guidelines should be crafted collaboratively, considering the inputs from all relevant stakeholders, and should evolve with advancements in technology and changes in societal expectations [17].

VII. LIMITATIONS

While advancing the application of LLMs in smart policing, this paper identified several limitations that warrant further exploration.

- The performance of the LLMs, including GPT-3 and GPT-4, demonstrated significant variability between the San Francisco and Los Angeles datasets. This underscores the challenge of adapting LLMs to datasets with different characteristics and features and diverse urban settings with unique crime patterns.
- Another important point and possible improvement of this research is performing an explicit calibration of the LLM outputs to achieve reliable probability estimates. For sensitive applications such as predictive policing, calibration is critical to ensure that the model's predictions can be interpreted accurately and confidently by law enforcement. This area has not been explored, representing a gap between model development and its practical, real-world application.
- Furthermore, although this paper acknowledges the importance of addressing potential biases and ethical concerns when deploying AI for crime prediction and smart policing in general, it does not provide a detailed framework for identifying, mitigating, and monitoring these issues.

VIII. CONCLUSION AND FUTURE WORK

The experiments in this paper demonstrate the feasibility of using LLMs in smart policing, with their superiority over traditional ML models in crime incident prediction tasks. The exploration of crime prediction models applied to datasets from SF and LA offers an in-depth understanding of the capabilities of LLMs, including GPT-3, and GPT-4, highlighting their capacity for crime classification. In particular, our fine-tuned GPT-3 model exhibited superior performance compared with traditional models, such as RF, in the SF dataset. However, the performance of LLMs on the LA dataset presents challenges and disparities, suggesting that their adaptability may vary according to the dataset characteristics. Interestingly, no single model has emerged as a powerful tool for the LA dataset. Although few-shot learning on GPT-4 exhibits marginal improvement, it still fails to comprehend the underlying complexities of the dataset. This result emphasizes the dataset-dependent nature of both the ML models and LLMs. The underlying reason for the disparities in the performance of the models on different datasets can be attributed to the different characteristics and features present in each dataset, and the struggles of the GPT models to adapt to some of these features in the LA dataset. This result emphasizes the dataset-dependent nature of both the traditional ML models and LLMs. Moreover, this indicates that crime prediction is a complex task as it depends on many factors related to an urban environment. Therefore, a comprehensive evaluation of different approaches is required to achieve the most suitable solution for this task before integrating them into real-world systems.

As LLMs have become integrated into broader applications of smart policing, there is growing scope for future work to address various challenges and explore opportunities. One significant avenue of exploration involves refining the capabilities of LLMs in analyzing recorded conversations between police and victims. Future research could focus on developing advanced techniques for extracting critical information from dialogues, improving sentiment analysis for an accurate understanding of emotional tones, and enhancing the recognition of relevant contexts to aid investigators in their decision-making process. In addition, the processing of videos and images from body cameras or security surveillance cameras presents a multifaceted challenge that future work can address. Research efforts may concentrate on refining computer vision algorithms integrated with LLMs to accurately recognize and track objects and individuals in real-time. Moreover, there is a need for advanced methods for event detection and anomaly identification within video streams to ensure efficient monitoring of complex urban environments. Future work in these domains should also prioritize ethical considerations, transparency, and accountability, ensuring the responsible deployment of these technologies in law enforcement to maintain public trust and safeguard individual rights. Our findings also emphasize the importance of conducting experiments in various scenarios to build more practical frameworks for smart policing. Although LLMs demonstrate superiority in certain contexts, their performance can be influenced by dataset-specific characteristics. Therefore, future research should explore the adaptability of language models to different urban environments by considering the interplay of socioeconomic, cultural, and geographical factors. Furthermore, the discrepancy between our results and related works underscores the complexity of evaluating models beyond their accuracy scores. Factors such as model interpretability, computational efficiency, and ethical considerations play a pivotal role in determining the most suitable model for real-world applications. A holistic evaluation that considers diverse metrics and practical implications is essential for ensuring the trustworthiness and scalability of predictive policing models. Future endeavors in the realm of smart policing and crime prediction should leverage these insights, emphasizing the context-specific nature of crime datasets and the continual evolution required in LLMs for optimal outcomes in diverse urban and cultural conditions.

ACKNOWLEDGMENT

The authors would like to thank Gary Bauer and Ian Bowles for their contributions to the research and discussions on smart policing.

REFERENCES

[1] S. S. Haghshenas, G. Guido, S. S. Haghshenas, and V. Astarita, "The role of artificial intelligence in managing emergencies and crises within smart cities," in *Proc. Int. Conf. Inf. Commun. Technol. Disaster Manage. (ICT-DM)*, Sep. 2023, pp. 1–5.

[2] H. Wang and S. Ma, "Preventing crimes against public health with artificial intelligence and machine learning capabilities," *Socio-Economic Planning Sci.*, vol. 80, Mar. 2022, Art. no. 101043.

[3] S. Maliphol and C. Hamilton, "Smart policing: Ethical issues & technology management of robocops," in *Proc. Portland Int. Conf. Manage. Eng. Technol. (PICMET)*, Aug. 2022, pp. 1–15.

[4] M. Afzal and P. Panagiotopoulos, "Smart policing: A critical review of the literature," in *Proc. 19th IFIP Int. Conf.*, 2020, pp. 59–70.

[5] L. Elluri, V. Mandalapu, and N. Roy, "Developing machine learning based predictive models for smart policing," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2019, pp. 198–204.

[6] M.-S. Baek, W. Park, J. Park, K.-H. Jang, and Y.-T. Lee, "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation," *IEEE Access*, vol. 9, pp. 131906–131915, 2021.

[7] X. Mu, "The platform construction of the traffic management smart police center under the background of 'Internet+,'" in *Proc. 5th Int. Conf. I-SMAC*, Nov. 2021, pp. 929–932.

[8] A. Jayakody, S. Lokuliyana, K. Dasanayaka, A. Iddamalgoda, I. Ganepola, and A. Dissanayake, "I-police—An intelligent policing system through public area surveillance," in *Proc. IEEE 12th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2021, pp. 0148–0154.

[9] F. Yang, *Predictive Policing*, Oxford Res. Encyclopedia, Criminology and Criminal Justice, Oxford Univ. Press, 2019, doi: 10.1093/acrefore/9780190264079.013.508.

[10] A. G. Ferguson, "Predictive policing and the future of reasonable suspicion," *SSRN Electron. J.*, p. 259, 2012.

[11] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[13] P. Sarzaeim, Q. H. Mahmoud, and A. Azim, "Experimental analysis of large language models in crime prediction," in *Proc. 37th Canadian Conf. Artif. Intell.*, 2024, pp. 1–19.

[14] P. Sarzaeim, Q. H. Mahmoud, A. Azim, G. Bauer, and I. Bowles, "A systematic review of using machine learning and natural language processing in smart policing," *Computers*, vol. 12, no. 12, p. 255, Dec. 2023.

[15] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Secur. J.*, vol. 21, nos. 1–2, pp. 4–28, Feb. 2008.

[16] T. Almanie, R. Mirza, and E. Lor, "Crime prediction based on crime types and using spatial and temporal criminal hotspots," 2015, *arXiv:1508.02050*.

[17] S. Raaijmakers, "Artificial intelligence for law enforcement: Challenges and opportunities," *IEEE Secur. Privacy*, vol. 17, no. 5, pp. 74–77, Sep. 2019.

[18] W. Safat, S. Asghar, and S. A. Gillani, "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," *IEEE Access*, vol. 9, pp. 70080–70094, 2021.

[19] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, "Improved method of classification algorithms for crime prediction," in *Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST)*, Aug. 2014, pp. 250–255.

[20] J. R. Khan, M. Saeed, F. A. Siddiqui, N. Mahmood, and Q. U. Arifeen, "Predictive policing: A machine learning approach to predict and control crimes in metropolitan cities," *Univ. Sindh J. Inf. Commun. Technol.*, pp. 17–26, 2019.

[21] R. Iqbal, "An experimental study of classification algorithms for crime prediction," *Indian J. Sci. Technol.*, vol. 6, no. 3, pp. 1–7, Mar. 2013.

[22] N. Ivan, E. Ahishakiye, E. O. Omulo, and R. Wario, "A performance analysis of business intelligence techniques on crime prediction," *Int. J. Comput. Inf. Technol.*, vol. 6, no. 2, pp. 84–90, 2017.

[23] E. Ahishakiye, D. Taremwa, E. O. Omulo, and I. Niyonzima, "Crime prediction using decision tree (J48) classification algorithm," *Int. J. Comput. Inf. Technol.*, vol. 6, no. 3, pp. 188–195, 2017.

[24] B. S. Aldossari, F. M. Alqahtani, N. S. Alshahrani, M. M. Alhammam, R. M. Alzamanan, N. Aslam, and Irfanullah, "A comparative study of decision tree and naive Bayes machine learning model for crime category prediction in Chicago," in *Proc. 6th Int. Conf. Comput. Data Eng.*, Jan. 2020, pp. 34–38.

[25] J. S. Shingleton, *Crime Trend Prediction Using Regression Models for Salinas, California*. Monterey, CA, USA: Naval Postgraduate School, 2012.

[26] T. Zia, M. S. Akram, M. S. Nawaz, B. Shahzad, A. Abdullatif, R. Mustafa, and M. I. Lali, "Identification of hatred speeches on Twitter," in *Proc. 52nd The IRES Int. Conf.*, 2022, pp. 27–32.

[27] M. Gayathri, "Suspicious activity detection and tracking through unmanned aerial vehicle using deep learning techniques," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 2812–2816, Jun. 2020.

[28] R. A. Berk, "Artificial intelligence, predictive policing, and risk assessment for law enforcement," *Annu. Rev. Criminology*, vol. 4, no. 1, pp. 209–237, Jan. 2021.

[29] A. Stec and D. Klabjan, "Forecasting crime with deep learning," 2018, *arXiv:1806.01486*.

[30] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 9, no. 3, pp. 139–154, Jun. 2016.

[31] A. Almehmadi, Z. Joudaki, and R. Jalali, "Language usage on Twitter predicts crime rates," in *Proc. 10th Int. Conf. Secur. Inf. Netw.*, Oct. 2017, pp. 307–310.

[32] *Gpt-4 Technical Report*, OpenAI, San Francisco, USA, 2023, *arXiv:2303.08774*, doi: 10.48550/arXiv.2303.08774.

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[34] S. Rezayi, Z. Liu, Z. Wu, C. Dhakal, B. Ge, C. Zhen, T. Liu, and S. Li, "AgriBERT: Knowledge-infused agricultural language models for matching food and nutrition," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 5150–5156.

[35] S. Yue, W. Chen, S. Wang, B. Li, C. Shen, S. Liu, Y. Zhou, Y. Xiao, S. Yun, X. Huang, and Z. Wei, "DISC-LawLLM: Fine-tuning large language models for intelligent legal services," 2023, *arXiv:2309.11325*.

[36] D. Liga and L. Robaldo, "Fine-tuning GPT-3 for legal rule classification," *Comput. Law Secur. Rev.*, vol. 51, Nov. 2023, Art. no. 105864.

[37] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, Jul. 2021.

[38] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," 2023, *arXiv:2303.14070*.

[39] J. Chen, X. Wang, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, J. Li, X. Wan, H. Li, and B. Wang, "HuatuoGPT-II, one-stage training for medical adaption of LLMs," 2023, *arXiv:2311.09774*.

[40] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.

[41] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "PIXIU: A large language model, instruction data and evaluation benchmark for finance," 2023, *arXiv:2306.05443*.

[42] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, Mar. 2019.

[43] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. M. J. Wu, "A review of generalized zero-shot learning methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4051–4070, Apr. 2023.

[44] W. Chen, "Large language models are few(1)-shot table reasoners," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 22199–22213.

[45] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[46] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020, *arXiv:2012.15723*.

[47] B. Zhang, H. Yang, and X.-Y. Liu, "Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models," 2023, *arXiv:2306.12659*.

[48] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," 2023, *arXiv:2308.10792*.

[49] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 27730–27744.

[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Jul. 2001.

[51] S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using XGboost," *Information*, vol. 9, no. 7, p. 149, Jun. 2018.

[52] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[53] DataSF. (2018). *Police Department Incident Reports: 2018 to Present*. [Online]. Available: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/about_data

[54] L. A. O. Dats. (2020). *Crime Data From 2020 to Present*. [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

PARIA SARZAEIM received the B.S. degree in computer science from Alzahra University, Iran, in 2021. She is currently pursuing the M.A.Sc. degree in electrical and computer engineering with Ontario Tech University, Canada. Her research interests include machine learning, generative artificial intelligence, and large language models.

QUSAY H. MAHMOUD (Senior Member, IEEE) was the Founding Chair at the Department of Electrical, Computer, and Software Engineering, Ontario Tech University, Canada, where he is currently a Professor of software engineering. He is also the Associate Dean of experiential learning and engineering outreach with the Faculty of Engineering and Applied Science. His research interests include intelligent software systems and cybersecurity.

AKRAMUL AZIM (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical, Computer, and Software Engineering, and the Head of the Real-Time Embedded Software (RTEMSOFT) Research Group, Ontario Tech University, Oshawa, ON, Canada. He is also a Professional Engineer in Ontario. His research interests include real-time systems, embedded software, software verification and validation, safety-critical software, and intelligent transportation systems.