Experimental Analysis of Large Language Models in Crime Classification and Prediction

Paria Sarzaeim*, Qusay H. Mahmoud, Akramul Azim
Department of Electrical, Computer, and Software Engineering Ontario Tech University, Oshawa, ON, L1G 0C5 Canada

Abstract

Increasing crime rates and evolving challenges in law enforcement have raised the need for innovative solutions, leading to the emergence of smart policing. This paradigm shift incorporates artificial intelligence (AI) with a specific focus on machine learning (ML) as a pivotal tool for data analysis, pattern recognition, and proactive crime forecasting. Large-language models (LLMs) as a subset of generative AI have been used in different domains, such as financial, medical, legal, and agricultural applications. However, the abilities and possibilities of adopting LLMs for smart policing applications such as crime classification remain unexplored. This paper explores the transformative potential of BART, GPT-3, and GPT-4, three state-of-the-art LLMs, in the domain of crime analysis and predictive policing. Utilizing diverse methods such as zero-shot prompting, few-shot prompting, and fine-tuning, this paper evaluates the performance of these models on state-of-the-art datasets from two major cities: San Francisco and Los Angeles. The goal is to demonstrate the adaptability of LLMs and their capacity to revolutionize conventional crime analysis practices. The paper also provide a comparative analysis of the aforementioned methods on the GPT series model and BART, in addition to ML techniques, showing that GPT models are more suitable for crime classification in most of our experimental scenarios.

Keywords: Large language models, LLMs, fine-tuning, zero-shot prompting, few-shot prompting, crime prediction

1. Introduction

The rise in crime rates and multiple challenges they pose have led to a growing demand for effective crime forecasting and prevention measures. Smart policing has emerged in this landscape in response to the pressing need for innovative solutions in law enforcement [1]. Police agencies, crime labs, and courts have increasingly embraced artificial intelligence for a wide range of purposes, such as administrative tools, facial recognition, surveillance cameras, DNA matching, and bail and sentencing. By harnessing the power of accumulated data and AI capabilities, smart policing offers a framework for data analysis and pattern recognition, enabling authorities to identify emerging criminal patterns and trends effectively [2]. These tools have the potential to speed up data processing and analysis for law enforcement while mitigating the influence of human biases [3].

Furthermore, predictive policing has been introduced as a research subfield of smart policing, which involves the use of 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, make forecasts regarding the occurrence of future crimes, and identify potential criminals and victims. Predictive policing leverages ML algorithms and statistical analysis methods to forecast criminal activities,

* Paria.Sarzaeim@ontariotechu.ca

including key details such as location, date, time, crime type, and potential victims, by analyzing both historical and real-time crime data [4]. 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 [5].

A large language model (LLM) is a class of artificial intelligence models, typically based on deep learning, specifically designed to comprehend and generate 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 [6]. 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 [7] and GPT-3 [6] have introduced a novel approach known as prompting. In prompting, there is typically no need for further training; instead, the model's input contains a task-specific text called a prompt. Prompt engineering involves creating, evaluating, and recommending prompts for natural language processing (NLP) tasks, enabling professionals to use LLMs for tasks such as data annotation, classification, and question answering. The other way to take advantage of powerful LLMs is by adopting them for specific tasks in fine-tuning.

The primary objective of this paper was to investigate the potential of LLMs within the domain of smart policing, with a specific focus on predictive policing using classification and prediction. Our approach involves the application of prompting and fine-tuning methods on BART and GPT models based on two state-of-the-art crime datasets of Los Angeles City and San Francisco City. 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 2 provides an overview of related work in crime classification and prediction, along with background information on LLMs. Section 3 discusses the prompting methods and fine-tuning methods used. Section 4 explains the structure of the datasets employed in the experiments. Section 5 presents the results collected from our experiments and a detailed discussion and comparison of the methods used. Finally, Section 6 concludes the paper and highlights potential avenues for future work.

2. Related Work

Predictive policing as a part of smart policing has been helpful in addressing crime-related challenges. These strategies rely on various techniques, including mapping techniques, ML, and NLP [8]. These techniques can be employed to classify crime incidents based on their characteristics and predict the future occurrence of crimes in specific locations, or to identify individuals who are most likely to engage in criminal activities.

Various mapping techniques have been employed to identify crime hotspots, which can be inferred as a basic form of crime prediction. These techniques include point mapping, thematic mapping of geographic areas, spatial ellipses, grid thematic mapping, and kernel density estimation (KDE). KDE is widely used for visualizing crime data because it effectively identifies hotspots without being constrained by geometric shapes such as ellipses [9]. Despite the popularity of these techniques, there are concerns about producing variations in maps with the same data or their potentially misleading apparel.

More recently, 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 [8]. A paper provides a comprehensive analysis of ML methods for crime prediction using crime datasets from Chicago and Los Angeles. This paper 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 [10]. In another paper, the authors used naïve Bayes and back propagation neural networks 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% [11], whereas, in a similar study on crime category prediction in the United States, decision trees outperformed naïve Bayes [12]. Decision tree is a popular and powerful ML algorithm in this context, as shown in [13], which compares decision tree j48 with a support vector machine, naïve Bayes, and neural network. The authors in [14] 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 algorithms were applied to the top nine selected features from the dataset, and as a result, the decision tree outperformed naïve Bayes.

Deep learning methods, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GAN), have gained attention in the realm of smart policing. 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 gained an accuracy of 74.1%, and their proposed method improved the accuracy by 1.5% [15].

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 [16]. Moreover, there is an emerging interest in utilizing NLP to analyze social media data, such as Twitter, to predict crime rates in cities [17].

In recent years, LLMs, such as GPT 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 [3]. The selection of BART, GPT-3, and GPT-4 for this paper was motivated by their distinct architectures and capabilities, which contributed to a comprehensive exploration of LLMs. Additionally, these models have been studied and considered for adoption across diverse applications, highlighting their versatility and potential effectiveness. GPT models offer a powerful contextual understanding, making them suitable candidates for tasks requiring 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 utilized 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 generalize effectively across various domains, making it proficient in tasks ranging from article writing to question answering and code creation [18]. GPT-4 was introduced in March 2023, representing a significant advancement by incorporating multimodal signal processing along 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 [19]. By including GPT-4 in our paper, we aimed to assess the impact of model advancements on crime prediction performance in comparison with GPT-3.

BART, short for bidirectional and autoregressive transformers, is another LLM with a denoising autoencoder designed for pre-training sequence-to-sequence models [20]. 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 [7] and GPT.

The capabilities of LLMs have been investigated across a range of domains, including agriculture [21], legal predictions [22], medical [23], and financial applications [24], utilizing various techniques for fine-tuning and prompting LLMs such as GPT-3 and BERT. However, the use of LLMs in crime analysis and classification remains unexplored. In this paper, we provide an experimental analysis 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 LLMs with that of the random forest ML algorithm as the baseline model for crime prediction.

3. Methodology

In this section, we present our comprehensive approach for exploration into the transformative potential of LLMs in crime analysis and predictive policing. Building on the paradigm of smart policing, we employ three state-of-the-art LLMs including BART, GPT-3, and GPT-4 by using fine-tuning and prompt engineering techniques such as zero-shot prompting and few-shot prompting.

The effectiveness of prompting LLMs relies on the quality and comprehensiveness of prompts as they guide the model's behavior. While simple prompts can be good for straightforward cases, providing additional information about the task and the desired output format can be beneficial to achieve optimal performance. The use of prompts is often associated with the manual engineering of prompts, where we have to carefully design and evaluate each element of the prompt to achieve the desired results in the crime prediction task. Our investigation on the application of zero-shot and few-shot prompting techniques to BART and GPT models allows us to explore their adaptability in predicting crime categories beyond their explicit training data.

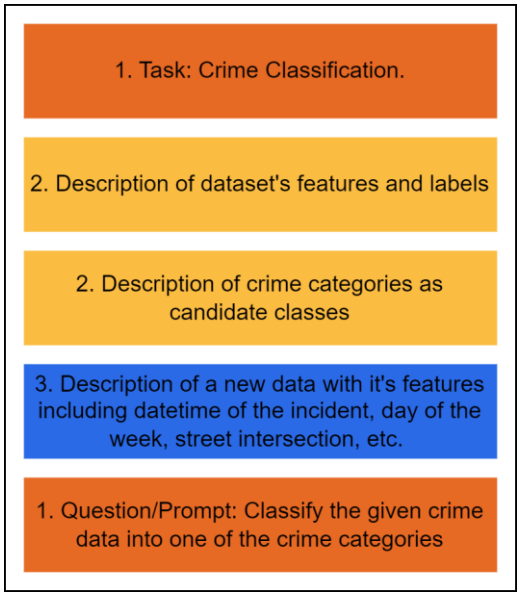

Figure 1. Prompt template that we used for zero-shot prompting and few-shot prompting

In the realm of ML, **zero-shot learning** means that a model is trained to recognize and classify instances that it has never seen before. This is typically achieved 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, enabling systems to generalize knowledge and make predictions beyond their initial training data [25].

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 was not explicitly trained for. This process involves employing text prompts that guide the model in understanding its tasks, the dataset, 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 [18].

For prompting in a zero-shot manner, we considered different templates to make the data understandable for LLMs. Eventually, we used a prompt template, as shown in Figure 1. Part (1) of the template specifies the task description for the model. Part (2) contains information about the dataset, such as its features and crime categories as candidate classes which should be listed and clarified in details for the model. Finally, part (3) provides the features of the new data that the model should classify. These features include the crime incident date, time, day of week, description, street intersection, and other related features to the location where the incident took place.

Few-shot learning in LLMs involves an approach similar to zero-shot learning, which 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 learning, K examples of input data and outcomes are provided, followed by one new input, and the model is expected to generate the appropriate outcome [18]. In this case, examples include the characteristics of a crime incident and the corresponding crime category of that incident.

The construction of prompts is important for the effectiveness of prompt-based learning approaches in both zero-shot and few-shot fashions. 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 by 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 [26]. The design of such prompts is similar to question answering prompts. We used the same template in Figure 1 for few-shot prompting LLMs, in addition to two examples of each class.

Fine-tuning is a technique in machine learning, rooted in transfer learning, where a model initially developed for one task is adapted for a related task. Pre-trained models, having learned features from extensive datasets, significantly expedite this process, proving particularly beneficial when labeled data is scarce or costly to obtain. Fine-tuning finds applications in domain adaptation, enhancing a model's performance on target data with different distributions, and data augmentation, where transformations on existing data create new samples to improve model performance, especially with limited labeled data. Fine-tuning language models refer to the adaptation of a new specific task by further training a pre-trained model on that task or its relevant dataset. This process often involves updating the model's weights and parameters based on a small task-specific data, while leveraging the general language understanding and generation capabilities learned during the pre-training phase. Instruction fine-tuning, also known as supervised fine-tuning, is identified as a critical process for acquiring instruction following the capability of LLMs and enhancing the capabilities and controllability of LLMs, as recognized in various real-world scenarios. This fine-tuning method entails refining pre-trained models on a dataset containing high-quality prompt-response pairs, aiming to align the model's behavior with human instructions rather than the typical next-word prediction objective of language models [27, 28].

The construction of instruction datasets 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. In this paper, we fine-tuned GPT models on an instruction dataset based on the original dataset of crime incidents in Los Angeles and San Francisco to classify crime incidents based on incident category.

For our experiemnts, we used the random forest (RF) machine learning model as the baseline. The RF is an ensemble learning model made up of several separate decision trees, each with a distinct set of attributes chosen at random. The final decision is made by the ensemble model by combining the output of these individual tree classifiers via a voting process. Random selection of split characteristics in each internal node and training sample selection using the bagging method [29]. One of the advantages of RF is its ability to to handle imbalanced datasets such as crime datasets that have an unequal number of instances across classes is one of its advantages. Predictive accuracy can be more meaningfully evaluated when the cost ratio between false positives and false negatives can be distinguished well. This model takes into account real-world consequences of false negatives and false positives and addresses the imbalances in the outcome variables, which increases the model's practical utility.

4. Experimental Results

To perform our experiments, we used two popular datasets for crime prediction, and in all scenarios of prompting or fine-tuning, we used the related OpenAI API for GPT models and the HuggingFace API to interact with the BART model. To reduce the complexity of the experiments, only the default settings for each model have been used to perform our experiments. For example, the temperature and top-p parameters of GPT models are set to 1 while the max number of output tokens is 64.

The first dataset includes more than 750,000 reports of crime incidents in San Francisco (SF) from 2018 to the present [30]. This dataset has 27 features and 41 unique incident types, which we used for the incident classification. We mapped these categories into a more manageable set of classes during the pre-processing phase. These categories serve as the labels on which the data should be classified. The graphical representation in Figure 2 (a) illustrates the uneven distribution of data across incident type categories with Property Crimes being the most prevalent crime.

The second dataset contains information on crime reports in Los Angeles (LA) from 2020 to the present [31]. This dataset includes more than 850,000 records of crime occurrences in LA City with 28 features. We selected 13 relative features and similar to the SF dataset, we set the crime categories as labels, and transformed them into a broader set of categories. Figure 2 (b) visually represents the data distribution of the LA dataset with Property Crimes and Violent Crimes being the most prominent in LA.

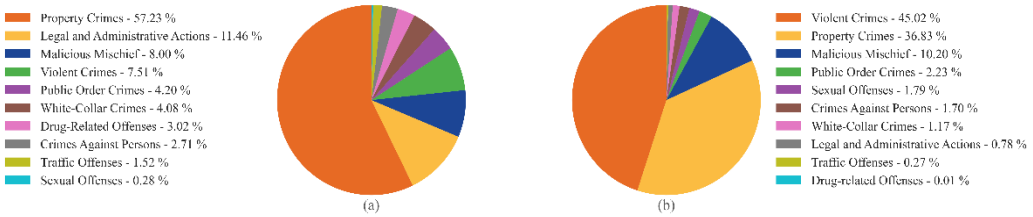

Figure 2. Data distribution; (a) Data distribution for the SF dataset, (b) Data distribution for the LA dataset

To evaluate the performance of the model in each scenario, we split the dataset into training and test sets; therefore, we used 80% for training and 20% for testing. It is also important to note that we used the same training set to train the RF model and fine-tune GPT-3; therefore, we had to reduce them in size according to the limitation and cost of fine-tuning

GPT-3 on the entire training dataset. As the datasets were imbalanced, we considered the weighted average of accuracy, precision, recall, and F1-score, as shown in Figures 3.

4.1. San Francisco Dataset

As shown in Figure 3(a), while the RF model as the baseline demonstrated strong performance and achieved a weighted accuracy of 82% on the SF dataset, the fine-Tuned GPT-3 outperformed all methods with an accuracy score of 97%. This result suggests that leveraging the contextual understanding of the pre-trained GPT-3 model, coupled with domain-specific fine-tuning, significantly improves its ability to distinguish different crime categories. Comparatively, the RF as a traditional supervised ML model, while achieving relatively good performance, fell short of the fine-tuned GPT-3 model, suggesting that the language model that is fine-tuned for crime classification provides a more effective method for the SF dataset. Additionally, zero-shot prompting on GPT-3 performed closely to the RF model, while BART exhibits the least favorable outcomes, and zero-shot prompting on GPT-4 outperforms both. This comparison shows that BART with its bidirectional architecture may offer advantages in capturing contextual dependencies, whereas GPT models, being autoregressive and pre-trained on extensive data, demonstrate significantly more powerful contextual understanding.

When compared to zero-shot prompting, the few-shot prompting approach has demonstrated modest improvement. This highlights the importance of giving the model precise samples of the data during prompting in order to improve its accuracy when classifying a new crime data. Furthermore zero-shot prompting GPT-4 showed a slight improvement over GPT-3, which can be attributed to the architectural and pretraining phase enhancements achieved in GPT-4. In general, whereas 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.

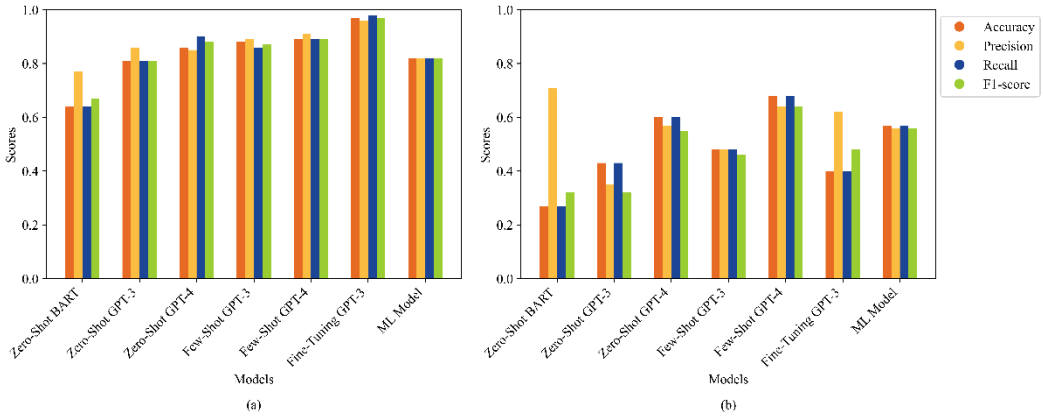

Figure 1. *Weighted Accuracy, precision, recall, and F1-score for different models; (a) Results for the SF dataset, (b) Results for the LA dataset*

4.2. Los Angeles Dataset

The results for the LA dataset reveal challenges in crime prediction, with notable differences compared with the SF dataset. According to Figure 3(b), the RF model as the baseline exhibited a weighted accuracy of 57%, suggesting a substantial drop compared to its counterpart in SF. Additionally, the fine-tuned GPT-3 achieved a weighted accuracy of 40%, which indicates a significant decrease in performance compared to its performance on the SF dataset. This discrepancy highlights the dataset-dependent nature of both the ML and LLMs models and emphasizes the need for approaches specific to characteristics of crime incidents in each location.

GPT-3 showed limited effectiveness, whereas GPT-4 achieved a comparatively higher accuracy of 60%. BART, on the other hand, continues to exhibit less favorable outcomes. Similarly, few-shot prompting contributed to the enhancement of performance for both GPT models, but the improvements were modest in comparison to the SF dataset. This suggests that few-shot learning still struggles to address the complexities and imbalances of crime occurrences within the LA dataset. Whereas incremental changes in GPT-4 do not necessarily translate into a clear advantage in this specific context, it can be a notable choice for crime prediction for this dataset. These results emphasize the need for careful evaluation and consideration of model capabilities for different tasks and datasets with different features.
Our experimental analysis's findings demonstrate that using LLMs in the realm of smart policing, specifically predictive policing, is feasible. While LLMs outperform traditional ML model in most scenarios, there are some circumstances with contradictory results, especially when using the LA dataset revealing that more experiments are required to build more practical solutions. Additionally, the comparison of our results with the related work shows that our Fine-Tuned GPT-3 has an impressive weighted accuracy surpassing all the other methods which is evidence of the potential of LLMs to outperform traditional ML techniques in crime classification and prediction. In the LA dataset, few-shot learning contributes to modest improvements but remains insufficient to overcome the complexities of crime occurrences in this specific urban environment.

However, it is crucial to recognize that the choice of LLMs and efficacy of fine-tuning depend on the nature of the dataset. The results on the LA dataset poses unique challenges, which are evident in the lower overall performance scores. An unexpected decrease in accuracy observed in our fine-tuned GPT-3 model on the LA dataset which revelas the importance of model selection based on crime data. 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 into 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.

Furthermore, ethical concerns regarding the use of such methods in predictive policing and their potential biases should be addressed. Potential biases in the training data can be utilized by these models, leading to unfair and discriminatory outcomes. It is essential to adopt strategies to identify and mitigate biases, ensuring that the deployment of LLMs aligns with the ethical standards. This includes ongoing monitoring and transparency in model decision making. 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 aims to define guidelines for transparency, comprehensibility, and ethical considerations. These guidelines should be crafted collaboratively, considering inputs from all relevant stakeholders, and should evolve with advancements in technology and changes in societal expectations [8].

5. Conclusion and Future Work

The experiments in this paper demonstrate the feasibility of using LLMs in crime prediction, with their superiority over traditional ML models applied to datasets from SF and LA highlighting the capabilities of LLMs, including BART, GPT-3, and GPT-4 in understanding and analyzing crime data. In particular, our fine-tuned GPT-3 model exhibited superior performance compared to traditional models, such as RF, in the SF dataset. However, the performance of LLMs on the LA dataset presents challenges and disparities. Interestingly, no single model has emerged as a powerful solution on the LA dataset, and although few-shot

learning on GPT-4 exhibits marginal improvement, it still fails to comprehend the underlying complexities of the dataset. Although LLMs demonstrate superiority in certain contexts, their performance can be influenced by dataset-specific characteristics. This result emphasizes the dataset-dependent nature of both ML and LLMs models in addition to the importance of conducting experiments in various scenarios to build more practical frameworks for crime prediction. Therefore, future research should explore the adaptability of LLMs to different urban environments, considering the interplay of socioeconomic, cultural, and geographical factors.

Furthermore, the discrepancy between the results 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 to ensure 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.

References

[1] S. Maliphol and C. Hamilton, "Smart Policing: Ethical Issues & Technology Management of Robocops," in *2022 Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland 2022: IEEE, pp. 1-15.

[2] J. Gable Cino, "Deploying the secret police: the use of algorithms in the criminal justice system," *Georgia Stet University Law Review*, vol. 34, 2018.

[3] A. Dixon and D. Birks, "Improving policing with natural language processing," in *Proceedings of the 1st Workshop on NLP for Positive Impact*, 2021, pp. 115-124.

[4] F. Yang, "Predictive policing," in *Oxford Research Encyclopedia of Criminology and Criminal Justice*, 2019.

[5] A. G. Ferguson, "Predictive policing and reasonable suspicion," *Emory LJ*, vol. 62, p. 259, 2012.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

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

[8] 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, 2023.

[9] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security journal*, vol. 21, pp. 4-28, 2008.

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

[11] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, "Improved method of classification algorithms for crime prediction," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014: IEEE, pp. 250-255.

[12] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. S. Panahy, and N. Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction," *Indian Journal of Science and Technology*, vol. 6, no. 3, pp. 1-7, 2013.

[13] E. Ahishakiye, D. Taremwa, E. O. Omulo, and I. Niyonzima, "Crime prediction using decision tree (J48) classification algorithm," *International Journal of Computer and Information Technology*, vol. 6, no. 3, pp. 188-195, 2017.

[14] 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 *Proceedings of 2020 6th International Conference on Computing and Data Engineering*, Sanya, China, 2020: Association for Computing Machinery, pp. 34-38.

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

[16] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 3, pp. 139-154, 2016.

[17] A. Almehmadi, Z. Joudaki, and R. Jalali, "Language usage on Twitter predicts crime rates," in *Proceedings of the 10th International Conference on Security of Information and Networks*, 2017, pp. 307-310.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and others, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877--1901, 2020.

[19] OpenAI, "GPT-4 Technical Report," 2023.

[20] 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," *arXiv preprint arXiv:1910.13461*, 2019.

[21] 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 *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, vol. 7, pp. 5150--5156.

[22] 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, 2021.

[23] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You, "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge," *arXiv preprint arXiv:2303.14070*, 2023.

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

[25] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

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

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, and others, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022.

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

[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[30] DataSF. "Police Department Incident Reports: 2018 to Present." https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/about_data (accessed December 20, 2023).

[31] L. A. O. Data. "Crime Data from 2020 to Present." https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data (accessed December 20, 2023).