iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76891

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

Privacy-Preserving and Trustworthy Fine-Tuning of Large Language Models

Mahammad Sadhik Karimuddin[1], Dr. K. V. Ramana[2], K. Ravi Kiran[3]
[1]*Student*, [2]*Professor*, [3]*Assistant Professor(c), Department of Computer Science and Engineering, JNTUK, Kakinada, India*

Abstract: ***Fine-tuning of large language models has been a common approach for adjusting these models for specific domains of their intended applications, but still, there are numerous issues related to the safe use of these models. Fine-tuning these models with sensitive or private data can result in the leakage of private data from these models in attacks such as membership inference attacks, gradient leakage attacks, or model inversion attacks. In addition to privacy-related concerns, fine-tuned language models are also faced with trust-related concerns due to their potential for being or becoming racist, toxic, or unverifiable. Furthermore, the efficiency of fine-tuning these models for real-world applications is also faced with efficiency constraints due to the computational requirements of these models. These issues have been addressed in the current state of the art in different ways, but the solutions for the above-mentioned problems are still traditionally addressed individually in their respective domains. With the help of the above-mentioned gaps in the related works, the proposed work aims to introduce a complete framework for differential privacy, data augmentation for synthetic data, model compression, as well as trust evaluation for the fine-tuning of language models.***

Keywords: Large Language Models, Fine-Tuning, Data Privacy, Differential Privacy, Synthetic Data, Model Compression, Trustworthy AI.

I. INTRODUCTION

Large language models have emerged as essential architectures for many of the recent breakthroughs in Artificial Intelligence, chiefly due to their ability to process and produce text reminiscent of human language. Nowadays, these models are being used extensively in many application domains, namely dialog systems, content generation, medical decision support, educational software, as well as financial analytics platforms. In most cases, these models are not used outright but are usually fine-tuned with application-specific data to make them more effective for these applications. While fine-tuning these models can bring forth many improvements in terms of accuracy as well as relevance, there are several issues that need to be taken care of before these models can be used for applications with certitude. However, one significant concern related to the fine-tuning procedure is sensitive data protection. Because these models can be trained on private data, they pose a risk of holding on to the information, which can later be exposed. In fact, existing literature has shown that a number of attacks such as membership inference, gradient leakage, and model inversion can be applied to retrieve information about the data from which the trained model was derived. This is even more significant for applications such as healthcare, law, and finance, which involve highly sensitive information confidentiality.

Besides privacy-related concerns, there is still a challenge in the trustworthiness of fine-tuned language models. This is because language models trained on real-world data may mirror or even exacerbate existing biases in the datasets they were trained on, create content that is harmful or misleading, or produce results whose trustworthiness is hard to determine. Such aspects might breed distrust in users and a myriad of other ethical concerns in case these models are employed on a mass scale.

Another challenge with a potential impact on the effective deployment of LMs is efficiency. The size of LMs coupled with their computational complexity incurs high memory and energy expenses, making it costly to deploy LMs. To this end, some techniques such as pruning and model quantization have been suggested to minimize computational complexity. However, this might negatively impact model behavior and performance whenever such techniques are employed separately.

Even though a range of techniques have been proposed for ensuring privacy, trust, and efficiency, such as differential privacy, synthetic data, federated learning, and model compression, among others, the majority of the existing techniques only handle individual issues and do not address them simultaneously. The problem mentioned creates a void in the literature for a comprehensive approach encompassing all these requirements.

©IJRASET: All Rights are Reserved

To fill this void, this paper examines the current literature, identifies the main limitations of the current solutions, and develops a comprehensive approach involving differential privacy, synthetic data augmentation, model compression techniques, as well as trust assessments to facilitate secure, reliable, and efficient fine-tuning of large language models.

II. LITERATURE REVIEW

Recent years have seen increasing research interest in overcoming the challenges related to privacy, trust, and efficiency in fine-tuning large language models. One of the earliest directions on this pertains to differential privacy, wherein noise is injected into model updates so that the influence of each individual training sample is limited. Techniques like differentially private stochastic gradient descent have been able to guarantee formal privacy; however, these approaches generally suffer from degraded performances when applied to large-scale language models because of the high extent of accumulated noise. In order to avoid these issues, there have been semi-supervised frameworks like PATE, which stands for Private Aggregation of Teacher Ensembles, where a set of teacher models trained on disjoint datasets provide privatized labels to a student model. Though it was shown to be effective on classification tasks, their scalability to generative language modeling has been rather poor.

Research also involves the use of synthetic data generation as a privacy-enhanced alternative for direct data training. Work in areas such as differential privacy GANs, together with PATE-GAN, aims to create synthetic data samples that can mimic the original data distribution while preserving privacy. Data can now be shared without revealing the original data, although the problem of generating good-quality synthetic text data is currently a challenge for these models. Adding noise for data privacy protection may end up diminishing the diversity in the text data, as well as promoting data biases in the training data, making these useless for NLP tasks.

Model inversion attacks and gradient-based attacks have reinforced the potential issues associated with collaborative or distributed learning conditions. Experiments regarding deep leakage through gradients and its refined extensions have shown that one can reconstruct the private training data based on the learnt gradients, especially in the federated or distributed learning setting. Such results have vindicated the importance of enhancing defenses for better privacy protection. However, additional overhead costs may arise due to the imposition of those defenses.

Efficiency-wise, much effort has been devoted to minimizing the deployment cost for large models via model compression techniques in order to lower the cost associated with deploying these models. Methods such as GPTQ, activation-aware weight quantization, are used in post-training quantization, supporting low-bit inference without degrading model performance. Pruning algorithms such as SparseGPT, Wanda, also remove unnecessary parameters while retraining is minimal. Even though the above algorithms are efficient in enhancing model efficiency, their objectives do not encompass privacy issues, trust, or fairness, such as bias, that may be compromised because of over-compression of the model.

Recent studies on trustable AI have also indicated the need to look into the limitations of the state-of-the-art approaches to mitigate bias and filter out toxicities. These studies have indicated that fairness and robustness need to be addressed within the optimization framework instead of considering them as separate adjustments. However, trust-based solutions have not addressed these areas when it comes to privacy protection mechanisms.

In general, the literature shows that existing efforts have focused on treating privacy, trust, and efficiency independently of one another. They include differential privacy methods that ensure data protection at the cost of utility, synthetic data methods lacking guarantees in data fidelity, compression-focused methods that ensure deployment efficiency but without any guarantees of trust, and methodologies aimed at fostering trust with no integrations on realizing privacy defenses. These limitations thus provide the motivation for having an integrated framework-one that can address privacy preservation, trustworthiness, and deployment efficiency jointly in the fine-tuning process of large language models.

III. EXISTING MODELS

The existing literature about the fine-tuning of large language models has various options for models and methods that can solve different issues associated with privacy and efficiency. The first set of models is based on Differential Privacy. It involves adding noise to the model during training to ensure private information is protected from being discovered through the model. The stochastic gradient descent model and the PATE model are examples of methods based on Differential Privacy. They ensure mathematical privacy for users, but their performance can be impacted for larger generative models.

Another set of models that have already existed targets synthetic data-driven model training. These models involve the generation of artificial datasets to reduce the need to work directly with sensitive information.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Methods like DP-GAN and PATE-GAN aim to retain the characteristics of the raw data statistically while ensuring that there are no privacy concerns. While these models do not pose risks of direct leakage of information, it is difficult to produce quality synthetic text.

Federated models based on federated learning have been recently suggested for mitigating privacy worries by conseptualizing data for decentralized training. In these models, the process of learning is distributed among various clients, where only model updates will be communicated to a central server. For adding robustness to privacy, secure aggregation schemes, as well as trusted execution environments, can be seriously utilized. Federated models suffer from a communication cost and can easily fall under gradient inference attacks for privacy preservation.

In terms of efficiency, model compression targeted models are designed to minimize the computation and memory complexity of larger language models. While quantization methods GPTQ and activation-aware weight quantization are aimed at lowering the precision of computations to facilitate low-bit computations, methods such as SparseGPT and Wanda are focused on eliminating unnecessary parameters. Even though such models are efficient in terms of deployment, they overlook aspects of privacy preservation and related trust attributes such as fairness. To sum up, current methods cover the issues of privacy, trust, and efficiency independently. Though all individual models are capable of yielding fruitful results in their realms, being independent does not make them very effective in actual scenarios where all these issues are interrelated. The above-mentioned attempt clearly emphasizes the importance of an overall approach being capable of managing all issues associated with the deployment of LLMs together.

IV. PROPOSED FRAMEWORK

A framework for addressing the shortcomings of current approaches might be one that takes into account privacy preservation, trustworthiness, and efficiency in a combined fine-tuning process for the task of fine-tuning large language models. Unlike traditional approaches that pursue multiple goals concurrently, the framework could be envisioned as breaking down these goals into interrelated subcomponents, which are to be optimized collectively. In the framework, the goals of privacy preservation in the context of highly sensitive data used for model training, trustworthiness in the context of fair performance of the model, and efficiency with respect to the model's computational requirements could be minimized in a balanced manner for a proper trade-off between the goals of privacy preservation, reliability of the model, and efficiency of the process. The overall workflow of the proposed system is illustrated in Fig. 1.

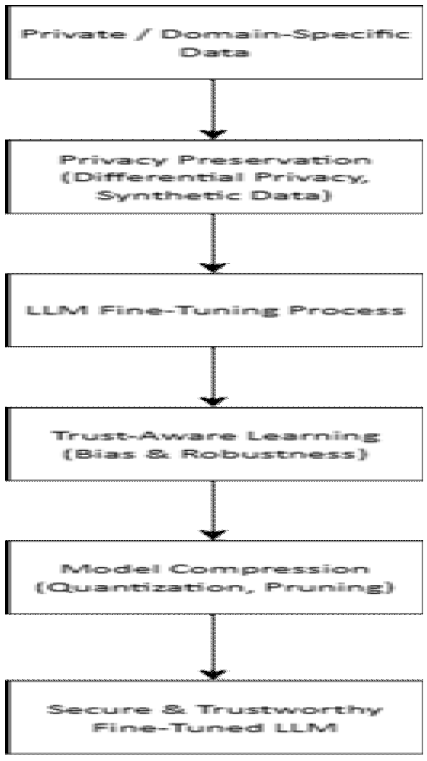


Fig. 1 Proposed Privacy-Preserving and Trustworthy Fine-Tuning Workflow

The workflow of the proposed framework is sequential and systematic, ensuring that the privacy of the users is maintained and the model is trustworthy and efficient during the process of fine-tuning the large language models. It starts with the selection of the training data that is domain-specific. It may involve some confidential data. However, to address the privacy issue, some mechanisms are applied that suppress the contribution of the individual data points. It makes it difficult to obtain the leaked data through the method of inference and the gradient.

In the subsequent stage, the technique of synthetic data augmentation is applied to complement the existing training data. Synthetic data that is protected from privacy violations is produced. Such synthetic data assists in the preservation of model utility and diversity without violating the privacy preservation constraint. The resulting dataset is then applied for fine-tuning the large language model.

After fine-tuning, there is a process flow that involves optimization techniques in order to enhance efficiency in deployment. This includes quantization methods that reduce precision to enable increased speed during inference. This is followed by pruning techniques that eliminate unnecessary parameters in order to further optimize efficiency in computation. This ensures that a fully fine-tuned model is obtained that is deployable even in environments where resources are limited.

Finally, the optimized model is then subjected to the trustworthiness evaluation phase to test its workings with respect to bias, robustness, and the production of toxic content. The findings of the study will be valuable to ensure that the privacy and compression strategies employed have not caused any unwanted consequences. If needed, modifications can be done to the earlier phases of the process to enhance the reliability. In this way, the proposed framework is capable of implementing the concept of secure and trustworthy optimization of the large language models.

V. RESULTS AND DISCUSSION

As the nature of the work is based on analytical results and conceptual framework, the findings shall be presented based on trend analysis and results of existing studies. Taking into consideration the results of the analysis of existing studies, the findings include the following:

Methodologies for privacy preservation such as differential privacy and PATE-based learning are successful in ensuring the security of leaked information but are associated with the reduced utility of the models.

Data generation and anonymization strategies are successful in protecting the models from direct exposure and access of sensitive information but fail to promote high-quality text generation.

Model compression strategies such as quantization and pruning are successful in improving the efficiency of the models but fail to address the concerns associated with trust and further have limitations regarding biased and robust models.

The proposed framework remedies the above deficiencies by bringing together privacy, trust issues, and efficiency strategies in a holistic pipeline. By leveraging differential privacy and the effectiveness of the ADAD method of synthetic data augmentation, the proposed framework minimizes reliance on private training data while preserving the efficiency of the model. The addition of quantization and pruning strategies ensures the efficiency of the fine-tuned model despite the differential privacy strategies. The proposed framework differs from the above said isolated approaches with the inclusion of trust evaluation strategies for identifying potential biases, robustness problems, and toxic content possibly caused by the above strategies.

A comparison between the state of the art and the proposed framework makes the advantages of this integrated approach salient. Current models usually optimize on a single objective, which results in having large trade-offs that perform worse in many aspects. Contrariwise, this proposed framework balances all objectives together: privacy risks are reduced while trustworthiness and efficiency are maintained. The proposed framework also does not pretend to totally remove all challenges associated with LLM fine-tuning; rather, it gives an organized and practical path toward how the concerns can be received in a co-operative manner.

In general, from this discussion, it can be seen that incorporating privacy protection, trust inference, and efficiency optimization into an integrated framework makes it possible to achieve more robust and practical large-scale language models. In contrast to previous models, this approach places equal stress on solving an entire problem or problem set, as opposed to aggregating multiple partial solutions to create an optimal solution.

VI. CONCLUSION

In this paper, the major challenges in fine-tuning large language models have been analyzed, with a special focus on privacy preservation, trustworthiness, and efficiency in model deployment.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

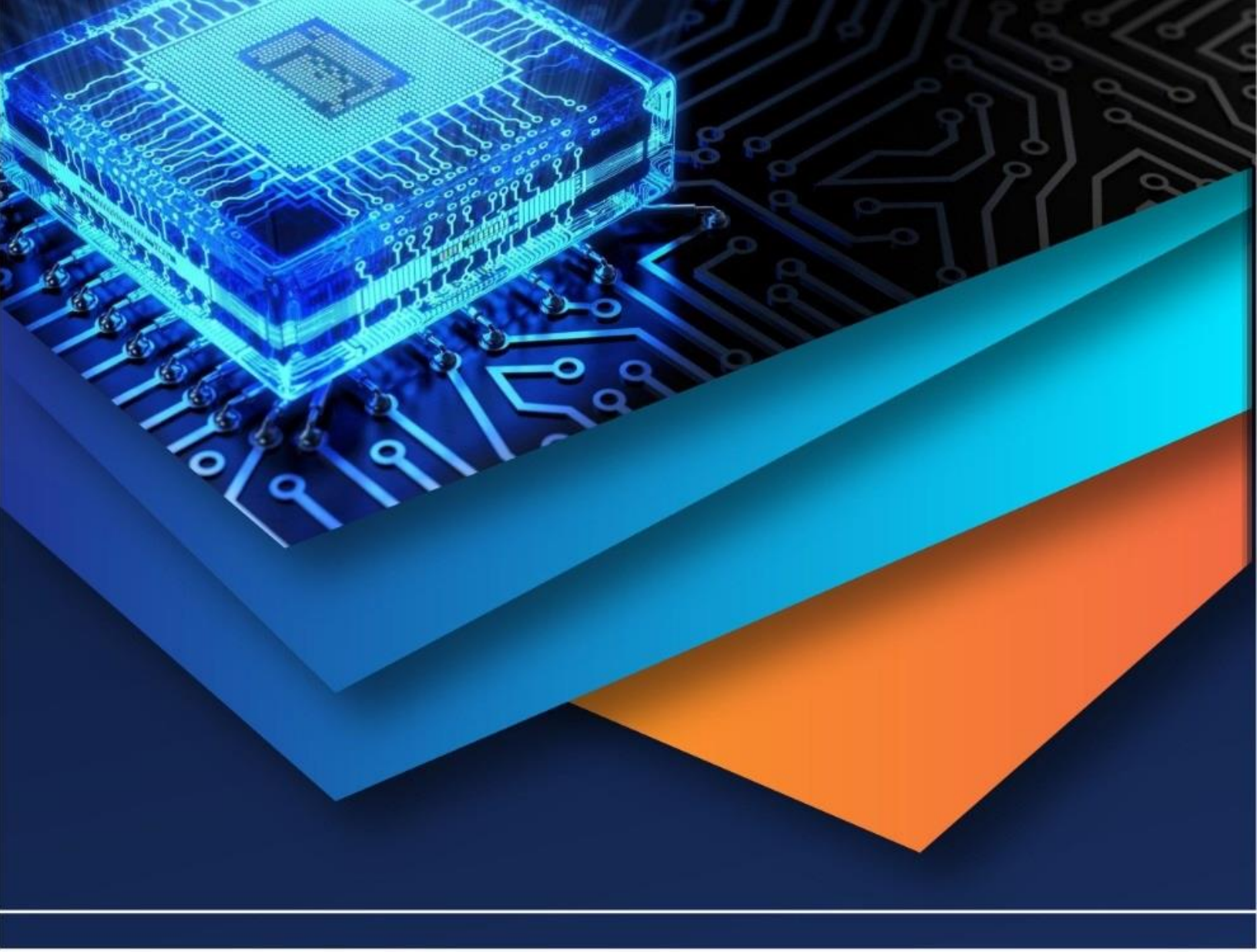
By giving a critical overview of the current literature, it has been found that most of these current models tackle one of these aspects at a time with some compromises in their performances in practical usage. Models focusing on privacy limit efficiency, efficiency-oriented methods ignore trust aspects, and trust-centric strategies hardly incorporate privacy mechanisms.

However, in order to remedy these shortcomings, this paper proposed an end-to-end framework that aims to combine privacy-friendly learning, data augmentation using synthetic data, model compression, and trust analysis in an overarching fine-tuning framework. By aligning the different aspects proposed in the new framework, it is ensured that there is an organized way of addressing challenges of data leakage, model trustworthiness, and efficient deployment while ensuring less impact on model performance. Even if the proposed framework is conceptual in nature, it has a vast potential in creating secure large language models.

The study, in general, points out the importance of jointly considering privacy, trust, and efficiency when finetuning large language models. This work contributes a holistic view that will guide future research and implementation efforts; this supports the wider adoption of large language models in sensitive and real-world domains where ethical considerations and resource constraints are pivotal.

REFERENCES

[1] A. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep Learning with Differential Privacy,” Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 308–318, 2016.
[2] N. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership Inference Attacks Against Machine Learning Models,” IEEE Symposium on Security and Privacy, pp. 3–18, 2017.
[3] B. Zhu, D. Han, and J. Park, “Deep Leakage from Gradients,” Advances in Neural Information Processing Systems (NeurIPS), pp. 14747–14756, 2019.
[4] M. Fredrikson, S. Jha, and T. Ristenpart, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,” ACM Conference on Computer and Communications Security (CCS), pp. 1322–1333, 2015.
[5] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, “Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data,” International Conference on Learning Representations (ICLR), 2017.
[6] J. Jordon, J. Yoon, and M. van der Schaar, “PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees,” International Conference on Learning Representations (ICLR), 2019.
[7] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, “Differentially Private Generative Adversarial Network,” arXiv preprint arXiv:1802.06739, 2018.
[8] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated Optimization: Distributed Optimization Beyond the Datacenter,” arXiv preprint arXiv:1511.03575, 2015.
[9] E. Frantar and D. Alistarh, “SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot,” arXiv preprint arXiv:2301.00774, 2023.
[10] J. Lin, J. Tang, H. Tang, S. Yang, and S. Han, “AWQ: Activation-Aware Weight Quantization for Large Language Model Compression,” arXiv preprint arXiv:2306.00978, 2023.
[11] J. Liu, R. Gong, X. Wei, Z. Dong, and J. Cai, “QLLM: Accurate and Efficient Low-Bitwidth Quantization for Large Language Models,” arXiv preprint arXiv:2310.08041, 2023.
[12] M. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” ACM Computing Surveys, vol. 55, no. 6, pp. 1–35, 2023.
[13] Z. Deng et al., “Hardening LLM Fine-Tuning: From Differentially Private Data Selection to Trustworthy Model Quantization,” IEEE Transactions on Information Forensics and Security, 2025.
[14] R. Carlini, S. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, and I. Goodfellow, “Extracting Training Data from Large Language Models,” USENIX Security Symposium, pp. 2633–2650, 2021.
[15] Y. Liu, X. Jia, H. Liu, and N. Z. Gong, “Privacy Risks of Fine-Tuning Large Language Models,” IEEE Symposium on Security and Privacy, pp. 213–230, 2022.
[16] J. Weidinger, L. Mellor, M. Rauh, C. Griffin, J. Uesato, P. Cheng, M. Glaese, R. McAleese, and I. Higgins, “Ethical and Social Risks of Harm from Language Models,” arXiv preprint arXiv:2112.04359, 2021.
[17] T. Brown, B. Mann, N. Ryder, et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), pp. 1877–1901, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)