Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

Towards Forgettable AI

[1]Swarwel Dorle, [2]Aditya karanjekar, [3]Yash Khawse, [4]Anand Uke, [5]Yash Wadbude, [6]Dr.Archana Dehankar

[1,2,3,4,5,6] Dept. Computer Technology Priyadarshini College of Engineering

Email: [1]swarwel.dorle@gmail.com, [2]adityakaranjekar@gmail.com, [3]yashkhawse13@gmail.com, [4]ananduke1@gmail.com, [5]yashwadbude2484@gmail.com, [6]archana.dehankar@pcenagpur.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Machine Unlearning (MU), Right to be Forgotten (RTBF), Data Privacy, GDPR

Abstract

This paper reviews the growing field of Machine Unlearning (MU), a discipline that supports modern privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), both of which provide users the right to have their data removed. Drawing on insights from recent research, the work presents an overview of MU principles, algorithms, and their implementation within artificial intelligence (AI) chatbot systems. The central problem addressed is how to efficiently remove the impact of specific data samples from trained models without retraining them entirely. Current literature presents two main categories: exact unlearning, which ensures complete data removal, and approximate unlearning, which offers faster and more practical performance. The techniques are broadly classified into data-based, model-based, and hybrid approaches. Federated Unlearning (FU), an extension of MU for distributed systems, introduces further complexities such as user participation and privacy leakage. Verification strategies, including adversarial evaluations like Membership Inference Attacks (MIA), are critical to validating that the system truly forgets deleted information. This paper concludes with recommendations for incorporating MU into conversational AI to promote transparency, privacy, and responsible data handling.

Introduction

The pervasive integration of machine learning (ML) models into daily life has brought significant concerns regarding data privacy and security. In response, legal frameworks such as GDPR and CCPA mandate that individuals have the right to request the erasure of their personal data, extending not only to databases but also to the influence of this data on trained models. Traditional approaches, such as retraining a model from scratch after a deletion request, are often impractical due to their high computational cost and time, especially for large, complex models. Machine unlearning addresses this challenge by providing efficient methods for selective data removal. Beyond regulatory compliance, machine unlearning enhances model robustness against poisoned or low-quality data, mitigates bias, and allows models to adapt to evolving datasets.As ML systems, including AI chatbots, become more complex and embedded in critical applications, unlearning becomes a key tool for building responsible, ethical, and trustworthy AI systems.

1. Foundational Concepts of Machine Unlearning Machine

Machine Unlearning is designed to make a model behave as if certain data were never part of its training process. The quality of unlearning is

assessed by comparing the unlearned model with a version retrained from scratch on the modified dataset. The two dominant paradigms are:
A. Exact Unlearning: Guarantees that the final model is statistically equivalent to a retrained version. It provides complete data removal but demands high computational cost.
B, Approximate Unlearning: Focuses on computational efficiency by achieving near-equivalence between unlearned and retrained models within acceptable margins.
Data deletion requests may involve single entries, classes, or features and can apply across continuous streams or dynamic model environments.The rapid growth of today's machine learning (ML) models has made their use essential in decision-making within finance, healthcare, and technology. However, this increased use raises serious issues about data security and retention. This situation has led to the rise of Machine Unlearning (MU), a new approach focused on developing methods to remove the influence of specific data that a model has learned. The main reason for this field's development is the many international data privacy laws, especially the General Data Protection Regulation (GDPR) and its important provision, the "Right to be Forgotten" (RTBF). Since ML models are unpredictable and may unintentionally retain the effects of training data even after it is deleted from the main database, it is vital to thoroughly "clean" them of this influence to comply with regulations. Besides meeting legal requirements, MU is important for protecting the integrity of models from data poisoning attacks and for getting rid of outdated or incorrect data in fast-changing settings like supply chains. In such environments, decisions must rely on the most current and accurate information.

2. Foundational Concepts and Theoretical Formulation

The main goal of machine unlearning can be seen as a reversed challenge. It aims to change a trained model ($\mathbf{M_o}$) into an unlearned model ($\mathbf{M_u}$) that looks and acts just like an oracle model ($\mathbf{M_r}$). The oracle model is trained only on the dataset that remains (D_r). The goal of being indistinguishable is key to the theoretical framework of the field. This framework categorizes solutions based on how much erasure is achieved. Exact Unlearning is the strictest form. It aims for complete mathematical similarity between the unlearned model and the oracle model. This provides the strongest legal guarantee. However, reaching this level of accuracy is usually not possible because it demands a lot of computing power to retrain complex models from the ground up. Therefore, much of the research focuses on Approximate Unlearning. This approach allows for a small, legally acceptable amount of influence to remain in exchange for significant efficiency gains. This compromise is essential when dealing with large-scale, enterprise systems. The nature of erasure requests also varies in specificity. These requests can range from Item/Sample Removal (removing individual data points) to Class Removal (excluding an entire category of output) or Feature/Attribute Removal (erasing specific traits like gender or race).

3. Methodologies for Data Erasure

The technical strategies for machine unlearning vary widely based on whether they rely on re-training or directly changing parameters. The main framework for effective Exact Unlearning is SISA (Sharded, Isolated, Sliced, and Aggregated). This method alters the traditional training process to lower the cost of future deletions. It starts by splitting the original dataset into several independent shards. Each shard trains a separate model, a process called isolation that limits the effect of any single data point. Additional efficiency comes from slicing. The training data for each shard is introduced incrementally. This allows the system to keep detailed checkpoints of model parameters. When a deletion request occurs, the system quickly identifies the specific affected model and resumes training from the exact checkpoint saved right before the data to be deleted was added. This leads to speed gains that make exact unlearning feasible.
In contrast, Approximate Unlearning techniques mainly work by directly altering the parameter space of the trained model. These methods often use tools like Influence Functions, a concept from robust statistics, which pinpoint the exact directional vector in the parameter space that indicates the influence of a specific data point. The core idea of the Certified Removal (CR) approach is to apply a "one-step Newton update" to this vector. It uses approximations of the complex Hessian matrix (such as the Fisher Information Matrix) to systematically eliminate the unwanted influence from the model weights. Other parameter-based methods include NegGrad (Gradient Ascent). This method aims to maximize the loss function on the deleted data, driving the model away from its previous memorization. However, it risks catastrophic forgetting of important retained knowledge. A simpler but often very effective alternative is basic Fine-Tuning. This simply continues training on the remaining data for a few epochs, using the

model's randomness to gradually reduce the impact of the deleted samples. Hybrid approaches combine these techniques within multi- objective frameworks. The Student-Teacher (S-T) paradigm uses knowledge distillation, where a student model is mentored by a Competent Teacher (CT), which embodies the retained knowledge, and an Incompetent Teacher (IT), which aims to introduce randomness for the forgotten data, to selectively curate information. A clear and creative approach is EntUn, which employs Entropy Maximization for the disregarded data. This method intentionally reduces the model's certainty in its predictions for the corrupted dataset, moving towards a random distribution. This process allows the model to "forget" the data in the output space while keeping the usefulness of the remaining data.

4. Applications Across Diverse Domains

Machine unlearning methods are increasingly tailored to address specific challenges in different fields. Federated Unlearning (FU) is an important area that deals with problems from decentralization. In this context, data isolation limits the central server's access to raw training data. Knowledge pervasion means that a client's influence has already spread among all other participants. Modern approaches for FU use Reinforcement Learning (RL) agents and Deep Q-Networks (DQN) to make flexible decisions about unlearning. They can choose dynamically between no erasure, partial erasure, or complete erasure based on a real-time balance of the client's assessed contribution, privacy costs, and computational needs. In Federated Recommendation systems, the CUFRU strategy handles these challenges by using a Gradient Transfer Station (GTS) on the server. This setup keeps track of historical model updates and enables iteration-aware gradient adjustments that effectively remove a departing client's influence from the overall model weights.

The area of Large Language Models (LLMs) presents a unique cost issue due to their large number of parameters, making traditional retraining very impractical. The DP2Unlearning framework addresses this issue by incorporating certified unlearning guarantees into the initial training phase. It uses Differential Privacy (DP) techniques like DP- SGD, which involves noise injection, or DP-MLM, which employs probabilistic token replacement.

This early safeguard allows for efficient fine-tuning on retained data to handle future deletion requests, providing a usable and certifiable method for removing LLM information.

For Generative Adversarial Networks (GANs), methods like Label Reversal force the model to forget the erased data quickly. This is done by changing its label from "real" to "fake" during training, disrupting the adversarial balance and effectively removing the sample's influence on the generator. Additionally, the DeltaBoost algorithm focuses on the sequential dependency challenges in Gradient Boosting Decision Trees (GBDTs) by introducing a solid architecture and innovations like gradient quantization to facilitate efficient updates when a deletion request is made.

5. Evaluation, Emerging Risks, and Future Research

Confirming that data has been completely erased is one of the most important unresolved issues in this field. This highlights the need for strong evaluation metrics. The most reliable measure is the Membership Inference Attack (MIA) Accuracy or Attack Success Rate (ASR). For successful unlearning, the model should reduce its ability to identify a deleted sample as a previous member of the training set to about $\mathbf{0.5}$, similar to random guessing. A broader measure, the Unlearning Effectiveness Score (UES), combines the ASR with a metric for Forgetting Quality, like the ZRF score. This aims for a unified target of $\mathbf{1.0}$. Beyond empirical validation, the ultimate goal is Certified Removal, supported by cryptographic "Proof of Unlearning" systems that external parties can audit to ensure compliance. However, the field continually faces new security threats that exploit the necessary steps of unlearning. Information- Stealing Attacks are particularly effective because they take advantage of the subtle differences in parameters and outputs between the original model ($\mathbf{M_o}$) and the unlearned model ($\mathbf{M_u}$) to infer membership or details about the deleted data. On the other hand, Model-Breaking Attacks seek to undermine the system's integrity. Malicious deletion requests can activate hidden backdoors in the updated model. These requests can also serve as Slow-Down Attacks that force costly full retraining by increasing the approximation error threshold of efficient unlearning techniques. Future research should focus on overcoming major technical challenges. This includes creating algorithms that can withstand Non-IID data deletion, which often leads to a significant drop in model performance. Additionally, effective unlearning solutions are needed for complex, time-sensitive architectures like Transformers and other cutting-edge models. Ultimately, it is crucial to continually improve auditable, robust, and scalable machine unlearning methods. This ensures that the

growing capabilities of artificial intelligence are matched by a strong commitment to user rights and data accountability.

The strategies for machine unlearning mainly fall into two categories: gradient manipulation and knowledge distillation. Gradient-based techniques, such as NegGrad (Gradient Ascent), aim to remove data by reversing standard optimization. Instead of focusing on minimizing the loss function, these algorithms try to increase the loss for the data to be forgotten. This approach shifts the model's parameters away from the learned state. While this method seems effective, it risks catastrophic forgetting. Aggressive changes to parameters can accidentally erase valuable knowledge related to the data that is still retained. In contrast, knowledge distillation frameworks, like the Student-Teacher (S-T) model in the FAST framework, handle this balance more carefully. In this method, a student model learns from a Competent Teacher, keeping important knowledge safe, and from an Incompetent Teacher, which encourages randomness or forgetting. This ensures the integrity of retained information during the erasure process. Another approach is EntUn, which tackles the "forget-retain dilemma" by using entropy maximization to reduce the model's prediction certainty about forgotten data to a random distribution. This effectively makes those samples indistinguishable from unseen data without destabilizing existing class boundaries.

The need for tailored unlearning strategies is especially clear in distributed and domain-specific learning systems, like Federated Learning (FL) and Learned Databases (DBs). Federated Unlearning (FU) faces challenges due to the inherent features of the FL framework. The data isolation principle prevents the central server from accessing the raw data needed for direct erasure. The principle of knowledge dissemination causes a client's impact to spread across the local models of other clients. To overcome these issues, solutions have emerged focusing on storing and manipulating the aggregated model updates. For instance, FL frameworks can use Reinforcement Learning (RL) agents, often with Deep Q- Networks (DQN), to assess each client’s status. They consider their contribution, privacy risks, and computational demands. This helps determine the best level of action, from total unlearning to partial unlearning, which helps maintain performance. On the other hand, Learned DBs adapt to environmental factors, with frequent updates of outdated data being common. Research shows that straightforward techniques like Fine- tuning and a mixed gradient approach called NegGrad+ can be effective in correcting learned components, such as cardinality estimators, after data deletion. This finding contrasts sharply with their high failure rates when trying to learn from updates based on new data insertion.

The rapid development of unlearning methods in specific, high-stakes areas highlights the need for tailored architectural solutions. For Large Language Models (LLMs), which are costly to retrain entirely, the best strategy involves adding Differential Privacy (DP) safeguards during the initial, resource-intensive training phase. Approaches like DP2Unlearning use DP techniques, such as gradient clipping (DP-SGD) or probabilistic token substitution (DP-MLM), to limit a data point's maximum impact on the model. This initial safeguard allows for efficient fine-tuning on the remaining data to meet subsequent erasure requests. This method provides a verified assurance of deletion at a much lower computational cost than naive retraining. Also, for deterministic, sequence-dependent models like Gradient Boosting Decision Trees (GBDTs), the DeltaBoost model introduces architectural resilience mechanisms. These include a self-balancing histogram and gradient quantization, which disrupt the dependency chain that would require cascading updates and retraining for each deletion. These advancements are crucial as they allow the unlearning process to be customized for both the data and the model's performance and structural weaknesses.

The future of machine unlearning will center on addressing issues with resilience and granularity, particularly in tough and complex environments. The challenge of deleting non-IID data continues to be a major hurdle. When the removed data samples are unevenly distributed, the unlearning process can disrupt model performance, leading to severe forgetting. Future research should focus on creating algorithms that can handle these distribution changes and on developing strategies to spot and fix the resulting data imbalances.

Moreover, there is a growing push for more detailed unlearning. This means moving away from broadly deleting whole samples or classes to enabling the exact removal of specific semantic features or data attributes from a model's latent space. This capability will be essential for managing complicated privacy requests and minimizing bias at a detailed level. Ultimately, the field needs to include strong cryptographic auditing and verification within these sophisticated algorithms to secure the whole compliance process. This ensures that technological progress remains closely tied to legal and ethical responsibility.

6. A Taxonomy of Unlearning Methodologies

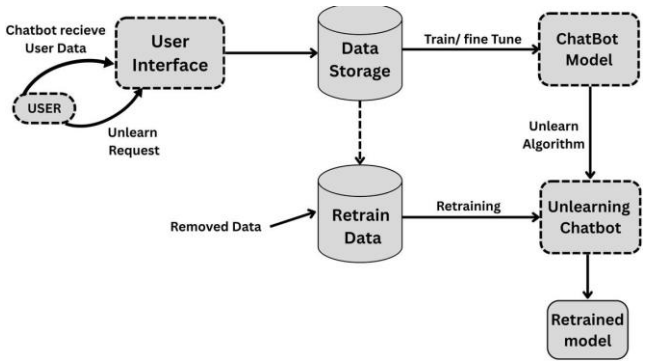
Machine unlearning techniques can be categorized based on whether they manipulate the training data, the model, or both. Data-oriented approaches focus on preprocessing or modifying the dataset to facilitate future unlearning. Data partitioning, exemplified by the SISA framework, divides the dataset into disjoint shards, training separate sub-models for each shard so that only affected sub- models need retraining upon a deletion request. GraphEraser extends this idea to graph data, retraining only the subgraphs influenced by the removed data. Data modification methods either convert algorithms into forms easily updated after deletions or inject error-maximizing noise to deliberately cause the model to forget specific classes. Model-oriented techniques operate directly on model parameters. Model reset approaches, such as DeltaGrad, reverse the learning process using gradient ascent on deleted data followed by fine-tuning, while Certified Removal uses single-step optimization for precise influence removal. Model modification methods, like DeltaBoost for gradient-boosted decision trees or Label Reversal for GANs, optimize architectures for efficient updates and minimize retraining overhead.

7. Federated Unlearning: A Specialized Domain

Federated learning (FL) trains models across multiple clients without central access to raw data. Federated unlearning (FU) applies MU principles to this decentralized environment, which introduces unique challenges. Knowledge permeation occurs when a client's data is diffused throughout the global model, complicating complete removal. Data isolation prevents direct access to client data, limiting traditional unlearning methods. Additionally, dynamic client participation, with frequent joining and leaving of clients, further complicates unlearning. FU strategies are categorized as passive, where the client leaves and the central server performs unlearning using historical data, and active, where the client participates in providing counteractive updates or fine- tuning. Reinforcement learning (RL) approaches, such as those using Deep Q-Network agents, dynamically determine whether to perform no, partial, or complete unlearning, balancing privacy, model performance, and computational efficiency.

8. Unlearning Verification, Attacks, and AI Chatbot Applications

Ensuring that data has genuinely been forgotten is critical. Verification often relies on adversarial attacks. Membership Inference Attacks (MIA) attempt to detect whether a specific data point was used during training, and effective unlearning should reduce an attacker's accuracy to random chance. Model inversion attacks aim to reconstruct removed data, while backdoor attacks inject hidden triggers that cause malicious outputs. Successful unlearning eliminates such vulnerabilities. Interestingly, the unlearning process itself can introduce privacy risks, as differences between the original and unlearned models may leak information about deleted data, a phenomenon known as information-stealing attacks.



For AI chatbots, machine unlearning directly addresses trust, privacy, and data security concerns. Applying model-oriented techniques to LLM architectures, minimizing differences between unlearned and original models, and implementing MIA-based verification mechanisms can provide auditable proof that a user's data has been forgotten, ensuring both utility and accountability.

Conclusion

Machine unlearning is a rapidly evolving field that reshapes how we approach data privacy and model accountability. By moving beyond full retraining, researchers have developed diverse methodologies— from data-partitioning

frameworks like SISA to model- specific techniques for GANs and GBDTs—that balance efficiency and privacy. The challenges become more complex in federated and distributed systems, requiring novel approaches. For AI chatbots, unlearning is a strategic tool for ethical AI, enabling selective data removal while maintaining performance. Implementing a robust, verifiable, and privacy-preserving unlearning framework tailored for conversational models, addressing secondary leakage, and providing concrete verification mechanisms demonstrates how AI can respect the "right to be forgotten" responsibly and effectively.

References

Afef Saihi, Mohamed Ben-Daya, Moncer Hariga, and Rami As'ad . A Structural equation modeling analysis of generative AI chatbots adoption among students and educators in higher education. Computers and Education: Artificial Intelligence, 7, 100274. DOI: 10.1016/j.cacai.2024.100274.(2024).

Bjørn Aslak Juliussen, Jon Petter Rui, and Dag Johansen. Algorithms that forget: Machine unlearning and the right to erasure. Computer Law & Security Review, 51, 105885. DOI:10.1016/j.clsr.2023.105885. (2023).

Chunxiao Li, Haipeng Jiang, Jiankang Chen, Yu Zhao, Shuxuan Fu, Fangming Jing, and Yu Guo. An overview of machine unlearning. High-Confidence Computing, 5(4), 100254. DOI:10.1016/j.hcc.2024.100254. (2025).

Dahuin Jung. EntUn: Mitigating the forget-retain dilemma in unlearning via entropy. ICT Express, 11, pp. 643-647. DOI: 10.1016/j.icte.2024.11.002.(2025).

Hoang Ngoc Tran, Nguyen Trung Nguyen, Nghi Vinh Nguyen, Ha Xuan Nguyen, and Anh Duy Nguyen. FAST: A pioneering unlearning framework integrating fine-tuning, adverse training, and student–teacher methods. Engineering Science and Technology, an International Journal, 64, 101996. DOI:10.1016/j.jestch.2025.101996. (2025).

Hengzhu Liu, Ping Xiong, Tianqing Zhu, and Philip S. Yu. A survey on machine unlearning: Techniques and new emerged privacy risks. Journal of Information Security and Applications, 90, 104010. DOI: 10.1016/j.jisa.2025.104010. (2025).

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine Unlearning: Solutions and Challenges. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(3), pp. 2150-2168.DOI: 10.1109/TETCI.2024.3379240. (2024).

Kongyang Chen, Dongping Zhang, Bing Mi, Yao Huang, and Zhipeng Li. Fast yet versatile machine unlearning for deep neural networks. Neural Networks, 190, 107648. DOI: 10.1016/j.neunet.2025.107648. (2025).

Kun Gao, Tianqing Zhu, Dayong Ye, Longxiang Gao, and Wanlei Zhou. Federated Unlearning With Reinforcement Learning: Adaptive Privacy Preservation for Clients. Journal of Information Security and Applications, 93, 104164. DOI:10.1016/j.jisa.2025.1041. (2025).

Lang Li, Pei-gen Ye, Zhengdao Li, Zuopeng Yang, and Zhenxin Zhang. Finetune and Label Reversal: Privacy- preserving unlearning strategies for GAN models in cloud computing. Computer Standards & Interfaces, 93, 103976. DOI: 10.1016/j.csi.2025.103976. (2025).

Lei Kang, Xuanshuo Fu, Lluis Gomez, Alicia Fornés, Ernest Valveny, and Dimosthenis Karatzas. Preserving privacy without compromising accuracy: Machine unlearning for handwritten text recognition. Pattern Recognition, 172, 112411. DOI:10.1016/j.patcog.2025.112411. (2026).

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In 2021 IEEE Symposium on Security and Privacy (SP), pp. 141-159. DOI: 10.1109/SP40001.2021.00019. (2021).

Jumde, A., Hazarika, I., & Akre, V. (2023). Challenges and opportunities in integrating rapidly changing technologies in business curriculum. In Proceedings of the 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 203–208). Dubai, United Arab Emirates: IEEE. https://doi.org/10.1109/ICCIKE58312.2023.10131683

Meghdad Kurmanji, Eleni Triantafillou, and Peter Triantafillou. Machine Unlearning in Learned Databases: An Experimental Analysis. Proc. ACM Manag. Data, 2(1), Article 49, pp. 1-26. DOI: 10.1145/3639304. (2023).

Sharma, B. (2025). Ethical and AI concerns in data privacy: A charismatic dilemma. International Journal of Multidisciplinary Research and Development, 12*(7)*, 18–32.

S. Schoepf, J. Foster, and A. Brintrup. (2025). Machine unlearning in supply chains. IFAC-PapersOnLine, 59(10), pp. 2945-2950. DOI: 10.1016/j.ifacol.2025.09.430. (2025).

Tamim Al Mahmud, Najeeb Jebreel, Josep Domingo- Ferrer, and David Sánchez. DP2Unlearning: An efficient and guaranteed unlearning framework for LLMs. Neural Networks, 192, 107879. DOI: 10.1016/j.neunet.2025.107879. (2025).

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A Survey of Machine Unlearning. ACM Transactions on Intelligent Systems and Technology, 15(1), Article 9, pp. 1-26. DOI: 10.1145/3629853. (2024).

Yang Li, Enyue Yang, Weike Pan, Qiang Yang, and Zhong Ming. Cross-User Federated Recommendation Unlearning. ACM Transactions on Intelligent Systems and Technology, 16(5), Article 119, pp. 1-24. DOI: 10.1145/3749990. (2025).

Youyang Qu, Xin Yuan, Ming Ding, Wei Ni, Thierry Rakotoarivelo, and David Smith. Learn to Unlearn: Insights into Machine Unlearning. IEEE Computer Magazine. (Specific reference details available upon request). (2024).

Zhaomin Wu, Junhui Zhu, Qinbin Li, and Bingsheng He. DeltaBoost: Gradient Boosting Decision Trees with Efficient Machine Unlearning. Proc. ACM Manag. Data, 1(2), Article 168, pp. 1-26. DOI: 10.1145/3589313. (2023).

Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, Xingliang Yuan, and Xiaoning Liu. A Survey on Federated Unlearning: Challenges, Methods, and Future Directions. ACM Computing Surveys, 57(1), Article 2, pp. 1-38. DOI: 10.1145/3679014. (2024).