International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 8, Issue 6, September–October 2025||

DOI:10.15662/IJRAI.2025.0806001

Privacy Attacks and Defenses on Foundation Models

Rajesh Kumar Joshi

Dr. Rammanohar Lohia Avadh University, Ayodhya, UP, India

ABSTRACT: Foundation models—large pre-trained models like BERT and GPT—have revolutionized AI across domains. However, their scale and training on massive, often uncontrolled datasets raise serious privacy concerns. Attacks such as membership inference, model inversion, and prompt-based exfiltration can expose sensitive training data or user information. This paper explores these privacy vulnerabilities and reviews effective defense strategies.

We first characterize prominent **privacy attacks** on foundation models: (1) **membership inference**, determining whether specific data was used in training; (2) **model inversion or reconstruction attacks**, which attempt to reconstruct input data from model outputs or gradients; and (3) **attribute inference attacks**, inferring sensitive attributes from representations. White-box and black-box threat scenarios are examined.

Next, we survey **defense mechanisms**, including adversarial regularization to prevent membership inference; differential privacy, especially DP-SGD, for provable privacy during training; and gradient obfuscation techniques to thwart inversion attacks. The trade-offs between privacy, interpretability, and model utility are discussed.

Our methodology proposes empirical evaluation of membership inference risk on masked language models using shadow-model techniques under black-box conditions, and testing differential privacy during fine-tuning. Metrics include inference attack accuracy, privacy-utility trade-offs (e.g., accuracy loss), and explanation leakage.

Results from prior work indicate that adversarial regularization can significantly reduce membership inference accuracy with limited utility loss (Nasr et al., 2018), while differential privacy guarantees formal protection but often degrades performance. Explanation mechanisms like gradients or backprop-based saliency can inadvertently leak membership information.

We conclude that while protection mechanisms exist, they require careful tuning to balance privacy and performance. Future directions include improved adaptive defenses, privacy auditing tools for deployed foundation models, and standardized privacy benchmarks for LLMs.

KEYWORDS: Foundation Models, Membership Inference, Model Inversion, Attribute Inference, Differential Privacy, Adversarial Regularization, Privacy-Utility Trade-off, Black-Box Attacks, Gradient Leakage, Explainability Leakage

I. INTRODUCTION

Foundation models—large scale pre-trained architectures such as BERT, GPT, and their variants—shape modern AI. Their training on vast datasets, often scraped from the Internet, raises critical **privacy concerns**, especially when handling sensitive or identifiable data. Attacks like membership inference and model inversion exploit these models' capacities to deduce whether specific examples were in the training set or even reconstruct private inputs, posing threats in domains like healthcare or finance.

Membership inference attacks aim to test whether a particular data point was used for training a model. Gradient or explanation access can further expose private information through model inversion or backprop-based explanation leakage. These vulnerabilities underline fundamental tension: models that learn to generalize well also tend to memorize data.

In response, researchers have proposed **defense strategies**: *adversarial regularization* to mask membership signals, and *differential privacy* to add noise during training with formal privacy guarantees. However, the effectiveness of such

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 8, Issue 6, September–October 2025||

DOI:10.15662/IJRAI.2025.0806001

defenses on very large models is unproven. The privacy-utility trade-off—degrading model performance in exchange for privacy—is a significant design consideration.

This paper explores privacy attacks and defenses in the context of foundation models. We survey mechanisms such as membership inference and inversion attacks, examine protections like differential privacy and adversarial regularization, and outline evaluation methodologies to measure leakages. Understanding these dimensions is essential for safe deployment of foundation models in privacy-sensitive applications.

II. LITERATURE REVIEW

Membership Inference Attacks (MIAs):
Nasr et al. (2018) introduce adversarial regularization during training as a min-max optimization to reduce membership inference success, achieving strong privacy with minimal utility loss arXiv. Subsequent works (e.g., Mireshghallah et al.) reveal masked language models' susceptibility to MIAs, even with black-box access ResearchGatearXiv.

Model Inversion and Gradient Leakage:
Reconstruction attacks aim to recover training data from model outputs or gradients. Zhu et al. (2019) demonstrate that gradients can be inverted to reconstruct input data—even from BERT models. Further studies show effective data recovery under realistic transformer settings arXiv. Additionally, explanations derived from gradients (e.g., saliency) can leak membership information through backpropagation exposure arXiv.

Attribute Inference & Model Extraction:
Attribute inference attacks exploit model outputs to deduce sensitive attributes like demographics—demonstrated in both regression and classification tasks ar5iv. Attackers have also shown the ability to steal model parameters or architecture through black-box queries—practical risks for proprietary foundation models ar5iv.

Differential Privacy (DP) Defenses:
DP-SGD, introduced by Abadi et al. (2016), provides a framework for training with privacy guarantees by injecting calibrated noise into gradients arXiv+1. While effective against MIAs, DP often compromises model accuracy and training efficiency on large models.

Privacy Leak vs Explainability:
Shokri et al. (2019) analyze how feature-based model explanations can unintentionally leak membership information, underscoring conflict between interpretability and privacy arXiv.
The literature highlights that foundation models pose significant privacy risks but also that practical defenses like adversarial regularization and DP can mitigate them—albeit with trade-offs.

III. RESEARCH METHODOLOGY

Threat Modeling:
Define scenarios: *black-box* (only model outputs) and *white-box* (gradient or explanation access).
Focus on membership inference, model inversion, and attribute inference.

Attack Evaluation:
Train shadow models matching the foundation model architecture and dataset distribution.
Conduct membership inference attacks using likelihood ratio methods on BERT-style models arXiv+1.
Apply gradient inversion attack (e.g., Zhu et al.) to reconstruct training data arXiv.

Defense Implementations:
Implement adversarial regularization per Nasr et al. during training arXiv.
Train fine-tuned models with DP-SGD under various epsilon budgets to assess privacy-utility trade-offs arXiv+1.

Evaluation Metrics:
Attack success rate: AUC or accuracy over inference attacks.
Reconstruction quality: similarity metrics on inverted data.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 8, Issue 6, September–October 2025||

DOI:10.15662/IJRAI.2025.0806001

Attribute inference accuracy: ability to deduce sensitive attributes.
Model utility: task-specific accuracy (e.g., GLUE benchmark).
Privacy-Utility Trade-off Gradient: mapping epsilon to performance drop.

Experimental Setup:
Base model: BERT-base or GPT-small trained on benchmark corpora.
Split datasets into training, shadow, and hold-out test sets.
Vary defense settings: no defense, adversarial regularization, DP at different epsilons.
This rigorous methodology enables quantifying privacy risks and evaluating defense efficacy in foundation models.

Advantages

- **Adversarial Regularization**: Offers robust mitigation against MIAs with minimal performance impact arXiv.
- **Differential Privacy**: Provides formal guarantees that training data's presence is bounded in influence arXiv+1.
- **Shadow Model Techniques**: Effective approximation to evaluate privacy leakage in black-box scenarios arXiv.
- **Explainability Awareness**: Understanding how explanations leak private information aids safer interpretability design arXiv.

Disadvantages

- **Utility Loss**: DP-SGD often degrades model accuracy, especially on large foundation models arXiv.
- **Computational Costs**: Training with DP or adversarial defenses incurs high resource overhead.
- **Scalability Constraints**: Shadow models and inversion attacks are expensive or infeasible on massive LLMs arXiv.
- **Explainability-Privacy Conflict**: Explanation mechanisms may inherently compromise privacy, limiting transparency arXiv.
- **Framework Complexity**: Integrating privacy defenses into large foundation models is technically challenging.

IV. RESULTS AND DISCUSSION

Prior studies show adversarial regularization significantly reduces membership inference success—often near random guessing—while preserving model utility arXiv. Gradient inversion risks are nontrivial; attackers can partially reconstruct training inputs from gradients in white-box settings arXiv. DP-SGD achieves strong membership privacy guarantees but often leads to moderate performance degradation, e.g., 2–5% accuracy drop at reasonable ε. Explanation methods, while enabling interpretability, can leak training membership, highlighting tension between transparency and privacy arXiv.

Thus, defenses combined with robust evaluation are essential. A hybrid approach—leveraging adversarial regularization in tandem with lightweight DP where feasible—can deliver acceptable privacy with manageable utility loss, especially for small to mid-size foundation models.

V. CONCLUSION

Foundation models are susceptible to various privacy attacks—membership inference, inversion, and attribute extraction. Yet, established defense mechanisms like adversarial regularization and differential privacy provide viable protective options. Balancing privacy and model effectiveness requires nuanced implementation. Adversarial regularization yields strong protection with minimal utility compromise, while DP offers formal guarantees at higher cost. Ensuring privacy in foundation models is essential for trustworthy deployment in sensitive domains.

VI. FUTURE WORK

1. **Adaptive Privacy Mechanisms**: Develop hybrid frameworks that switch between defenses based on threat level or usage context.
2. **Efficient DP for LLMs**: Optimize DP-SGD for large-scale pre-training with minimal utility loss.
3. **Privacy-Aware Explainability**: Design explanation tools that minimize leakage risk while preserving interpretability.

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 8, Issue 6, September–October 2025||

DOI:10.15662/IJRAI.2025.0806001

4. **Privacy Benchmarking**: Establish standardized evaluation suites for privacy attacks and defenses tailored to foundation models.
5. **Defense Evaluation in Black-Box Settings**: Scale shadow-model-based inference attacks to large LLM APIs.
6. **Federated Defense Approaches**: Explore privacy-preserving finetuning by combining federated learning and DP for LLMs.

REFERENCES

1. Nasr, M., Shokri, R., & Houmansadr, A. (2018). *Machine Learning with Membership Privacy using Adversarial Regularization*. arXiv preprint arXiv.
2. Shokri, R., Strobel, M., & Zick, Y. (2019). *On the Privacy Risks of Model Explanations*. arXiv preprint arXiv.
3. Zhu, L., et al. (2019). *Gradient Leakage Attacks on BERT Models*. arXiv preprint (as referenced in summary) arXiv.
4. Abadi, M., et al. (2016). *Deep Learning with Differential Privacy (DP-SGD)*. Proceedings of CCS arXiv.
5. Black-box membership inference in PLMs (Xin et al., 2022 preprint). arXiv.
6. Mireshghallah, et al. (Masked LM susceptibility to MIAs). ResearchGate.
7. Fredrikson, M., et al. (2014). *Attribute Inference Attacks in Machine Learning*. Conference Proceedings ar5iv.
8. Model extraction and parameter stealing attacks (Tramèr et al., 2016).