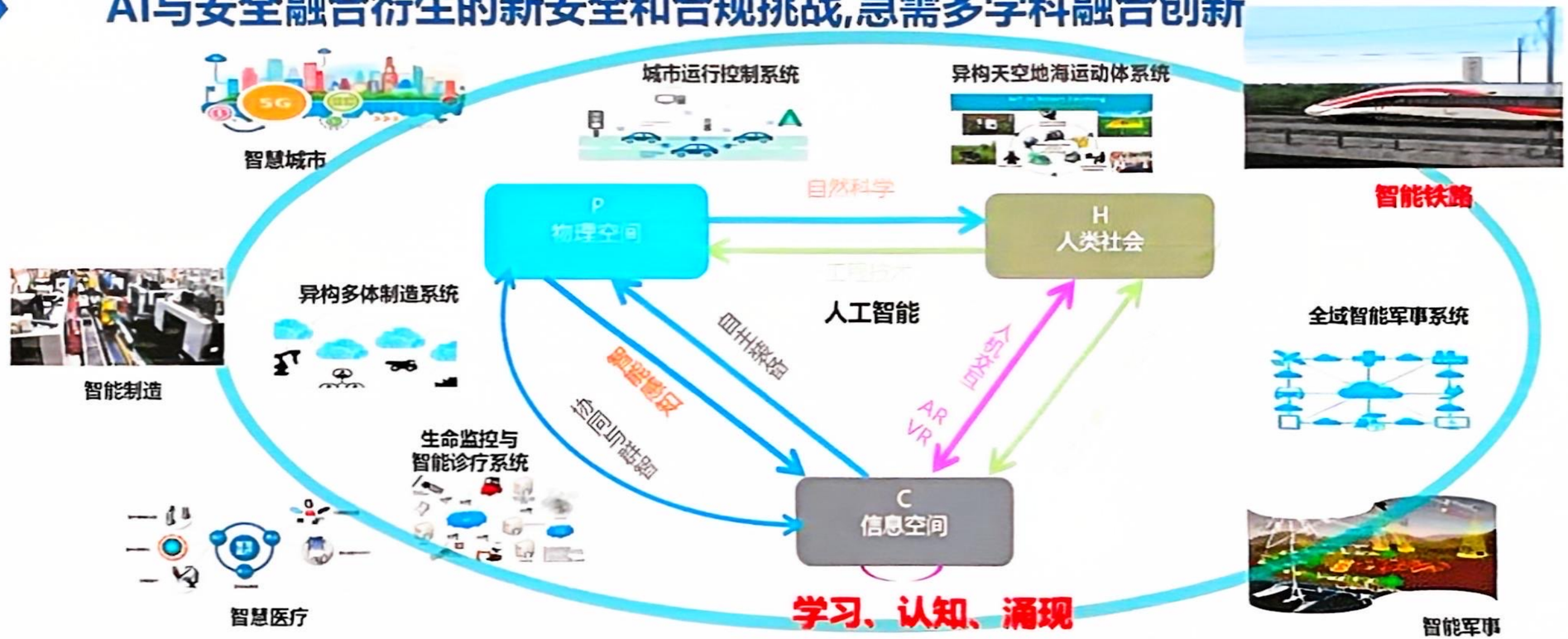
数字化智能化正在深刻地渗透进产业发展和社会治理基础设施，成为科技和法律必须协同面对的重大机遇和挑战

AI与安全融合衍生的新安全和合规挑战,急需多学科融合创新



- 人工智能全面赋能助力网络安全新突破
- 内生矛盾：AI的涌现性 vs 合规规制的相对滞后

AI的训练和演化，离不开大量的数据或语料，硬币的另一面：

内生问题： VS **法律合规等挑战**

训练过的模型会不会
不易察觉地
泄露隐私
或重要数据？

数据安全和隐私保护

重要数据的跨境合规

AI模型内生的隐私泄露风险难以察觉难以度量

难以对风险目标进行检索、识别

- 无可解释性的信息编码方式
- 内容护栏
- 其他对抗或隐藏措施

常规数据治理风险评估措施往往无从下手

Prof C X tan

AI隐私风险治理合规需求

法律法规对隐私信息保护

美国《加利福尼亚州消费者隐私法》在2024年的修订中明确规定个人信息可以以各种格式存在，包括人工智能系统，

欧盟2024年发布的《欧盟人工智能法》根据风险对人工智能进行分级为：不可接受风险，高风险，有限风险以及极小风险，

我国的《生成式人工智能服务管理暂行办法》中也强调了风险分级管理以及行业差异化监管的必要性，要求在算法设计、数据选择、模型优化等环节采取有效措施，以防止歧视及保护个人隐私权、名誉权等，

法律法规对数据流动监管

《网络安全法》第三十七条规定关键信息基础设施的运营者在中华人民共和国境内运营中收集和产生的个人信息和重要数据应当在境内存储，

《数据安全法》第三十一条规定关键信息基础设施的运营者在中华人民共和国境内运营中收集和产生的重要数据的出境安全管理，适用《网络安全法》的规定，

《个人信息保护法》第三十八条规定个人信息处理者因业务等需要，确需向中华人民共和国境外提供个人信息的，应当具备通过国家网信部门组织的安全评估、按照国家网信部门的规定经专业机构进行个人信息保护认证、按照国家网信部门制定的标准合同与境外接收方订立合同等条件之一，

AI数据安全治理相关法律及合规要求

法律依据

- **网络安全法** — 2017年6月1日起施行
- **数据安全法** — 2021年9月1日起施行
- **个人信息保护法** — 2021年11月1日起施行

法规依据

- **数据出境安全评估办法** — 2022年9月1日起施行
- **促进和规范数据跨境流动规定** — 2024年3月22日起施行
- **工业和信息化领域数据安全管理办法（试行）** — 2023年1月1日起施行

2024年9月29日发布

《工业和信息化领域数据安全合规指引（征求意见稿）》

有问题，有要求，[redacted]

合规需求

序号	类型	名称
1	法律	《中华人民共和国数据安全法》
2		《中华人民共和国网络安全法》
3		《中华人民共和国个人信息保护法》
4	部门规章	《促进和规范数据跨境流动规定》
5		《数据出境安全评估办法》
6		《数据出境安全评估申报指南（第二版）》
7		《个人信息出境标准合同办法》
8		《个人信息保护认证实施规则》
9		《工业和信息化领域数据安全管理办法（试行）》
10		《工业和信息化领域数据安全风险评估实施细则（试行）》
11		《工业和信息化领域数据安全事件应急预案（试行）》

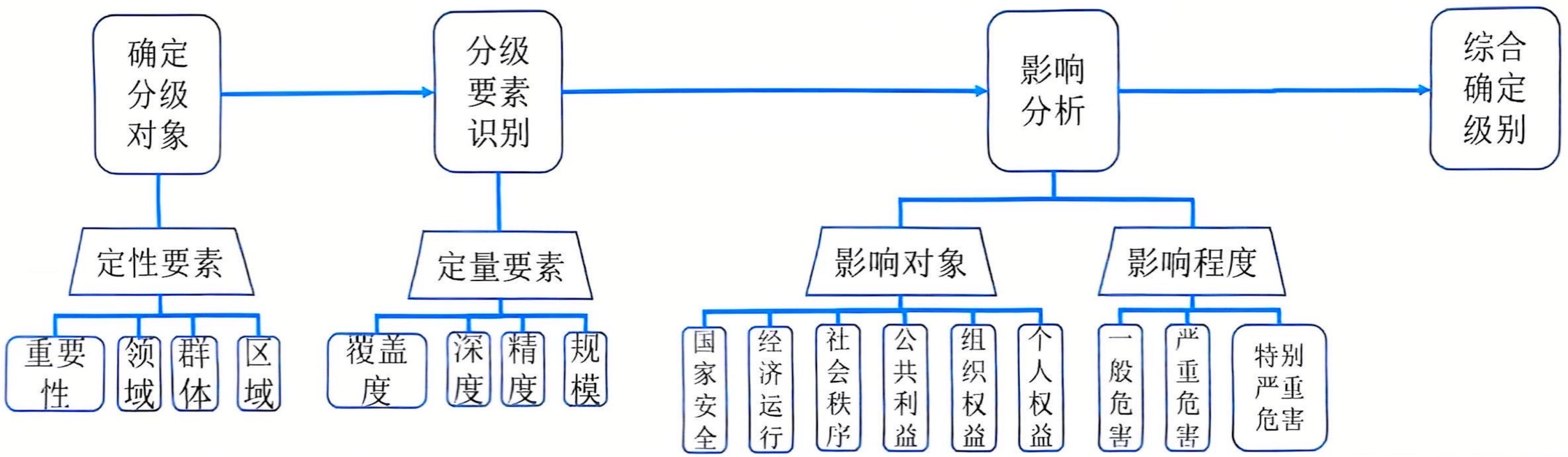
序号	类型	名称
12	标准指南	《工业领域重要数据识别指南》
13		《工业企业数据安全防护要求》
14		《工业领域数据安全风险评估规范》
15		《电信领域重要数据识别指南》
16		《电信领域数据安全分级防护要求》
17		《电信领域数据安全风险评估规范》

讨论主题：

1. AI模型内生的重要数据或隐私泄露风险治理的合规需求
2. **模型内生隐私数据泄露机理**
3. 典型科学难题与解决方案探索
4. 未来展望

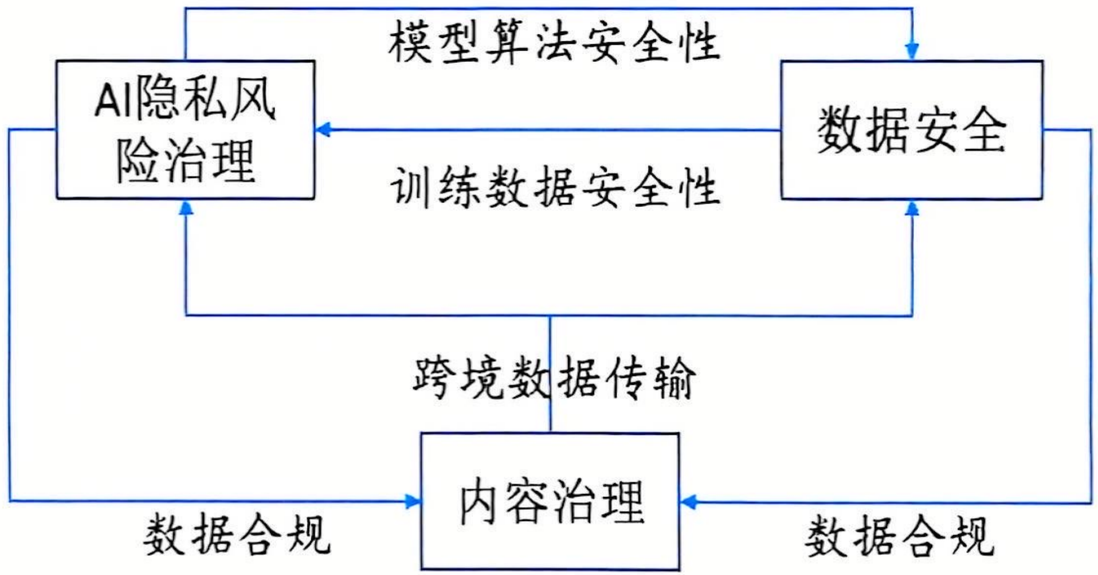
数据分级分类与隐私治理

数据分级分类制度在隐私风险治理中至关重要，因为它为不同敏感度的数据提供了差异化的保护措施，从而确保了个人隐私的安全性和合规性。GB/T 43697-2024《数据安全技术 数据分类分级的规则》给出了详细的数据分类分级的通用规则，为数据分类分级管理工作的落地执行提供重要指导。



AI隐私风险治理

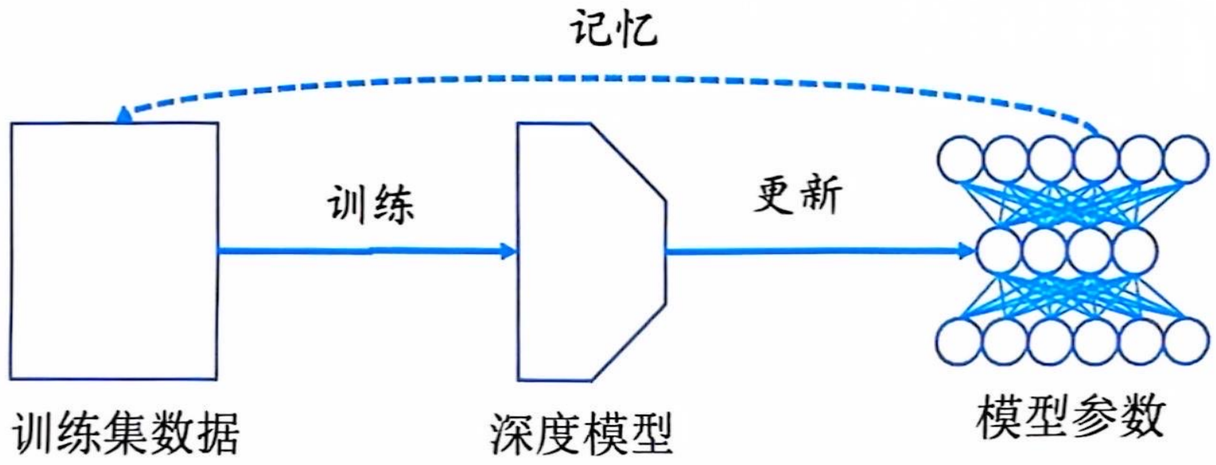
AI隐私风险治理、数据安全与内容治理之间存在着紧密的交叉关系。



然而，AI混合着公共数据和私人数据的大数据被挖掘、整合、分析、利用，人们难以凭自己的感官察觉到隐私权被侵害，个人信息安全面临着更高的风险。

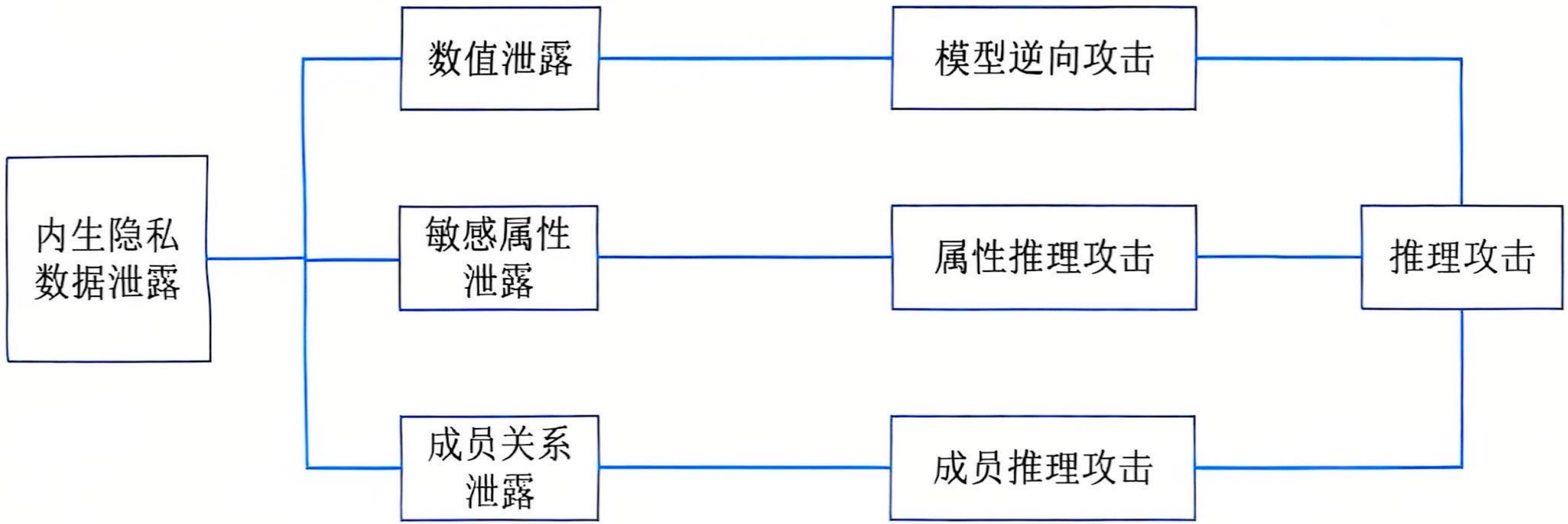
AI隐私风险治理

深度学习模型，尤其是具有大量隐藏层的深度神经网络，将某些数据的细节编码为模型参数，甚至记忆整个数据集，传统隐私保护的分级分类制度难以具体到每一个模型参数，训练集信息处理的合规性缺乏保障。



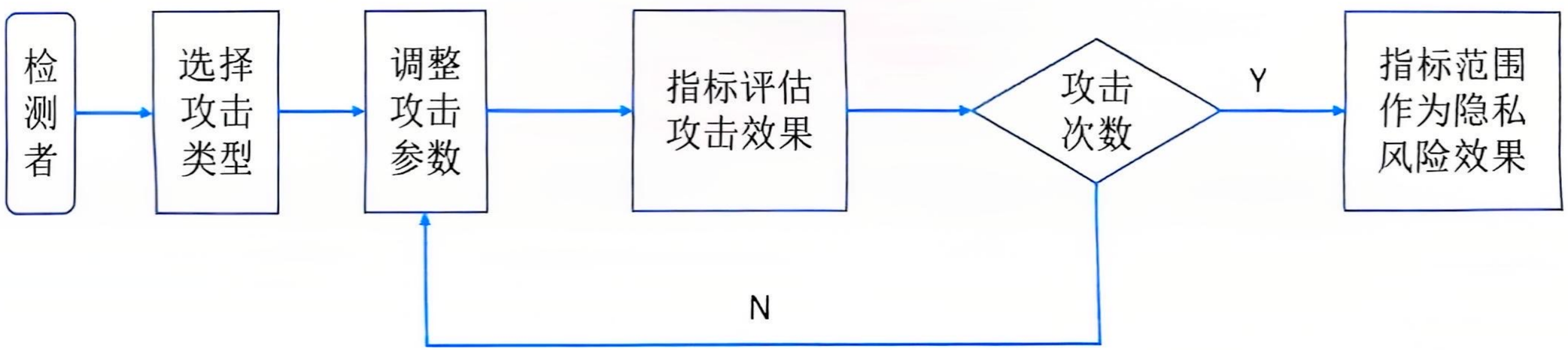
模型内生隐私数据泄露机理

内生隐私是指在使用人工智能模型时，训练集中的数据或者用户输入的包含隐私数据的信息被模型无意识地记录并存储，从而导致隐私泄漏的风险



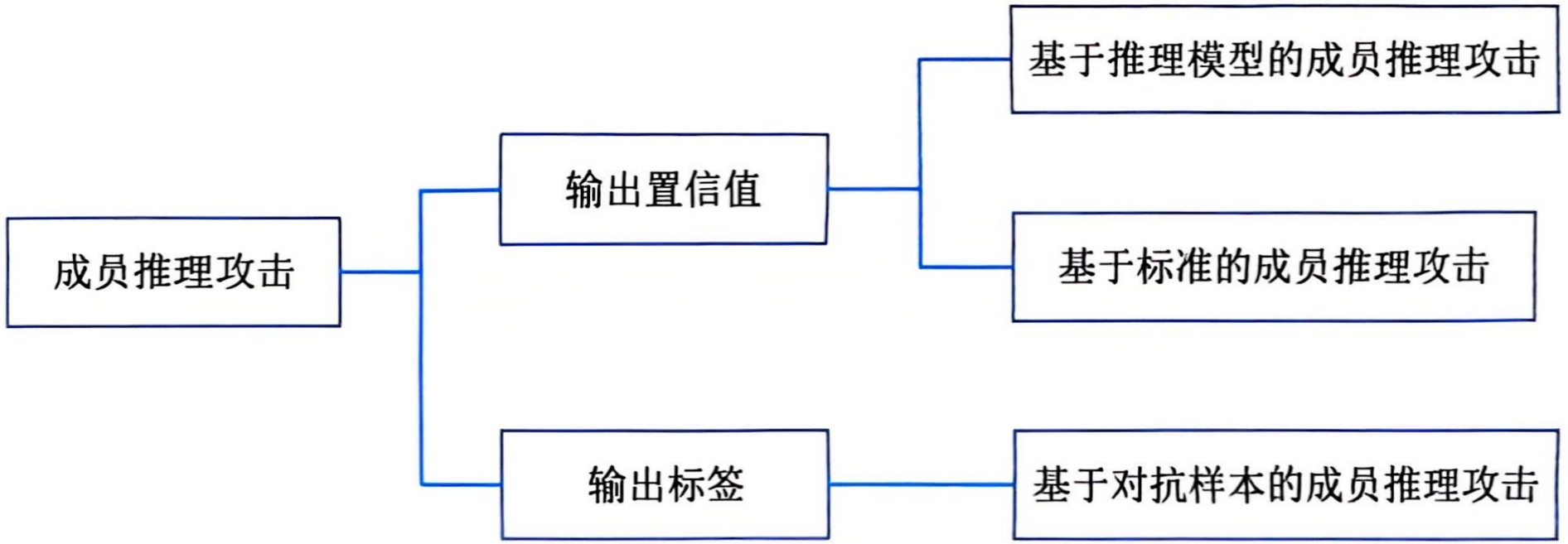
模型内生隐私数据泄露评估

评估方法	原理	优点	不足之处
还原检测	对受害者模型采用不同的攻击方法，攻击效果的指标作为评估指标量化造成的隐私风险。	简单而直观，能及时更新最近的攻击，时效性强	评估结果受到训练集、模型结构、训练集算法多种因素共同影响，难以泛化到一般模型



成员推理攻击相关研究

由于成员隐私代表了最基本的隐私泄露形式，在一定条件下可以转化为数值泄露与敏感属性泄露，更具有研究价值。



基于推理模型的成员推理攻击

攻击提出者	攻击原理	不足之处
Shokri[1]	每个类训练了一个影子模型，以模仿目标模型在该类中的行为，并根据影子模型中每个类的后验结果训练了神经网络	每个类都需要训练影子模型，数量过于庞大
Salem[2]	所有类只训练一个影子模型模仿目标模型	假阳率过高，容易把非成员识别为成员
Liu[3]	对目标模型知识蒸馏，将样本在不同迭代轮次下蒸馏模型的预测损失作为推理模型输入	需要保留大量蒸馏模型

[1] Shokri R , Stronati M , Song C ,et al.Membership Inference Attacks against Machine Learning Models[J].IEEE, 2017.DOI:10.1109/SP.2017.41.
[2]Salem A, Zhang Y, Humbert M, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models[J]. arXiv preprint arXiv:1806.01246, 2018.
[3] Liu Y, Zhao Z, Backes M, et al. Membership inference attacks by exploiting loss trajectory[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022: 2085-2098.

基于标准的成员推理攻击

攻击提出者	攻击原理	不足之处
Yeom[1]	成员样本的预测交叉熵小于非成员样本，通过预设的阈值进行区分	过于依赖模型过拟合
Song and Mittal[2]	通过考虑与类别相关的阈值以及地面实况标签，改进了基于熵的攻击	假阳率过高，容易把非成员识别为成员
Ye[3]	提出LOOD度量方法，也能识别信息泄漏的原因和高泄漏率的位置	使用高斯过程建模目标模型训练算法的随机性，对敌手要求权限高

[1] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]//2018 IEEE 31st computer security foundations symposium (CSF). IEEE, 2018: 268-282.
[2] Song L, Mittal P. Systematic evaluation of privacy risks of machine learning models[C]//30th USENIX Security Symposium (USENIX Security 21). 2021: 2615-2632.
[3] Ye J , Borovykh A , Hayou S ,et al.Leave-one-out Distinguishability in Machine Learning[J]. 2023.

基于对抗样本的成员推理攻击

攻击提出者	攻击原理	不足之处
Yeom[1]	成员样本预测成功率高，预测标签的样本即为成员样本	过于依赖模型过拟合
Choquette[2]	使用HSJA算法生成对抗样本，并且把对抗样本与样本之间的范数距离作为样本鲁棒性的量化指标，通过预设的阈值区分成员与非成员样本	假阳率过高，容易把非成员识别为成员
Chaudhari[3]	在模型训练过程中进行投毒使训练集样本在模型输出中具有特定特征，更容易识别	对敌手权限要求高

[1] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]//2018 IEEE 31st computer security foundations symposium (CSF). IEEE, 2018: 268-282.
[2] Christopher A. Choquette-Choo Florian Trame`r Nicholas Carlini Nicolas Papernot.1abe1 on1y membership inference attacks[C]//International Conference on Machine Learning.PMLR, 2021.
[3] Chaudhari H, Severi G, Oprea A, et al. Chameleon: Increasing Label-Only Membership Leakage with Adaptive Poisoning[J]. arXiv preprint arXiv:2310.03838, 2023.

讨论主题：

1. AI模型内生的重要数据或隐私泄露风险治理的合规需求
2. 模型内生隐私数据泄露机理
3. **典型科学难题与解决方案探索**
4. 未来展望

我们关注的主要科学难题

问题一：面向多敌手推理攻击的隐私泄露公平评估

现有的隐私风险评估局限于针对特定的推理攻击，而一个模型可能面临多种推理攻击。这些推理攻击前置条件不一，难以实现公平比较

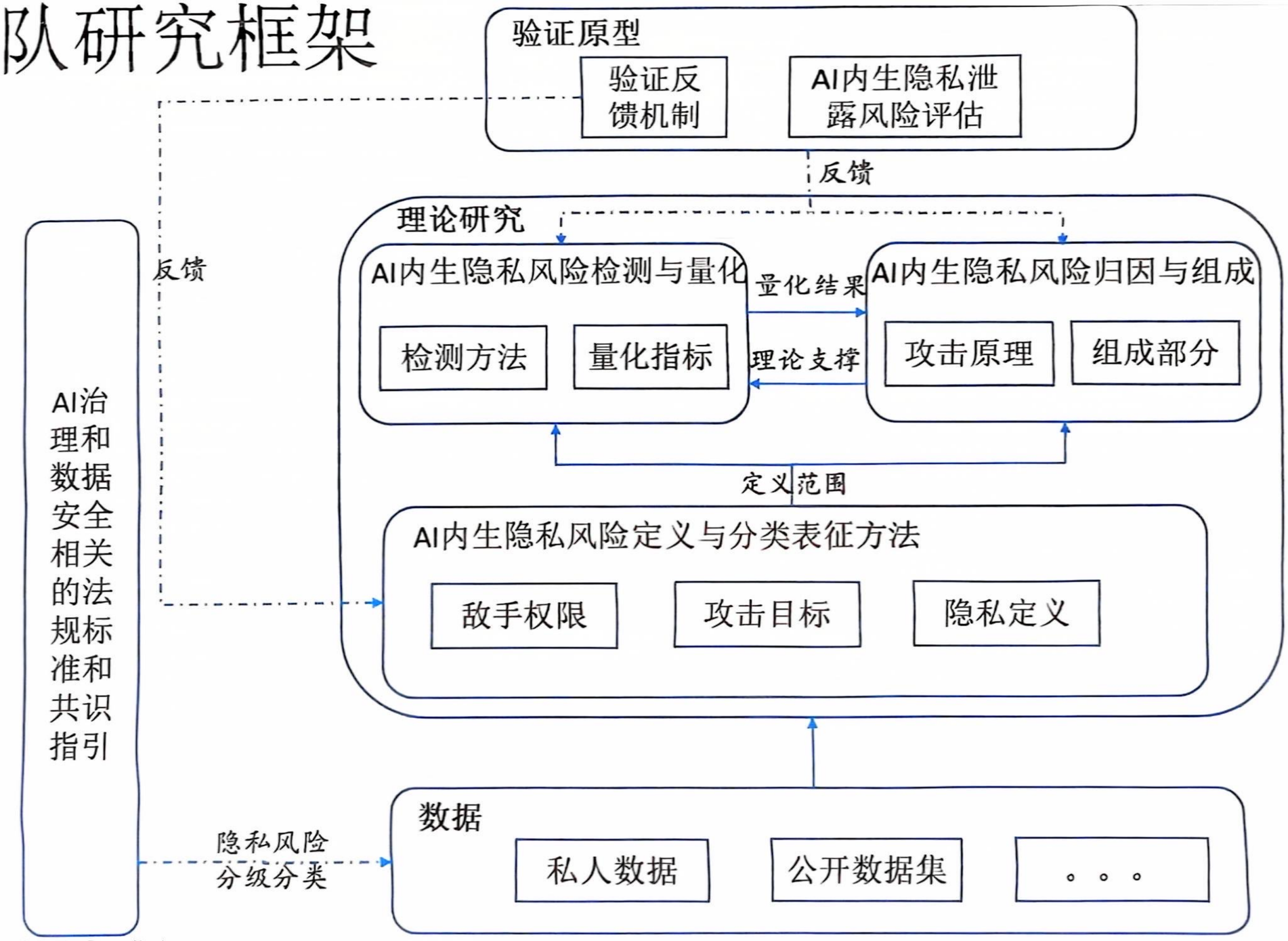
问题二：隐私泄露评估结果如何去除多因素影响

推理攻击隐私泄露评估结果往往受到目标模型训练集分布、目标模型结构、目标模型训练算法等多重因素影响，难以准确评估模型的隐私风险

问题三：合规取证理论基础

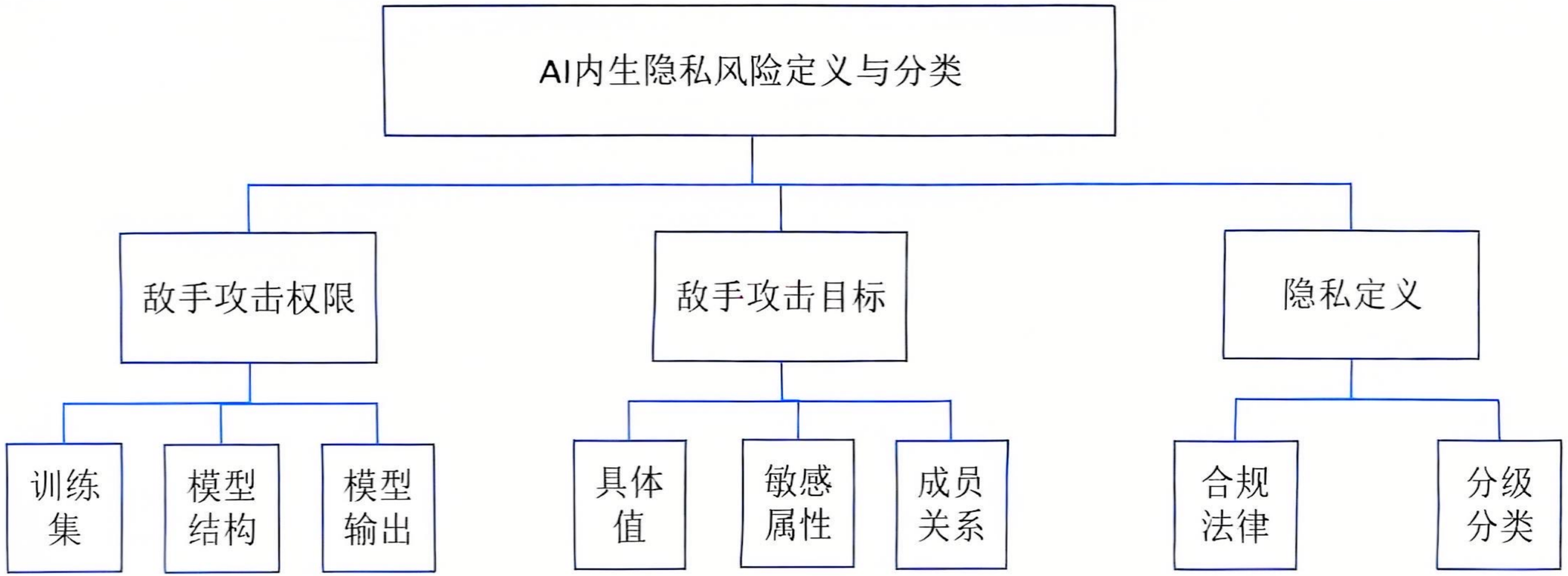
推理攻击往往基于直觉性假设，缺乏对应的理论依据支撑，训练集中不同类型的样本隐私泄露的可能性也不相同，也缺少隐私泄露的合规取证

团队研究框架

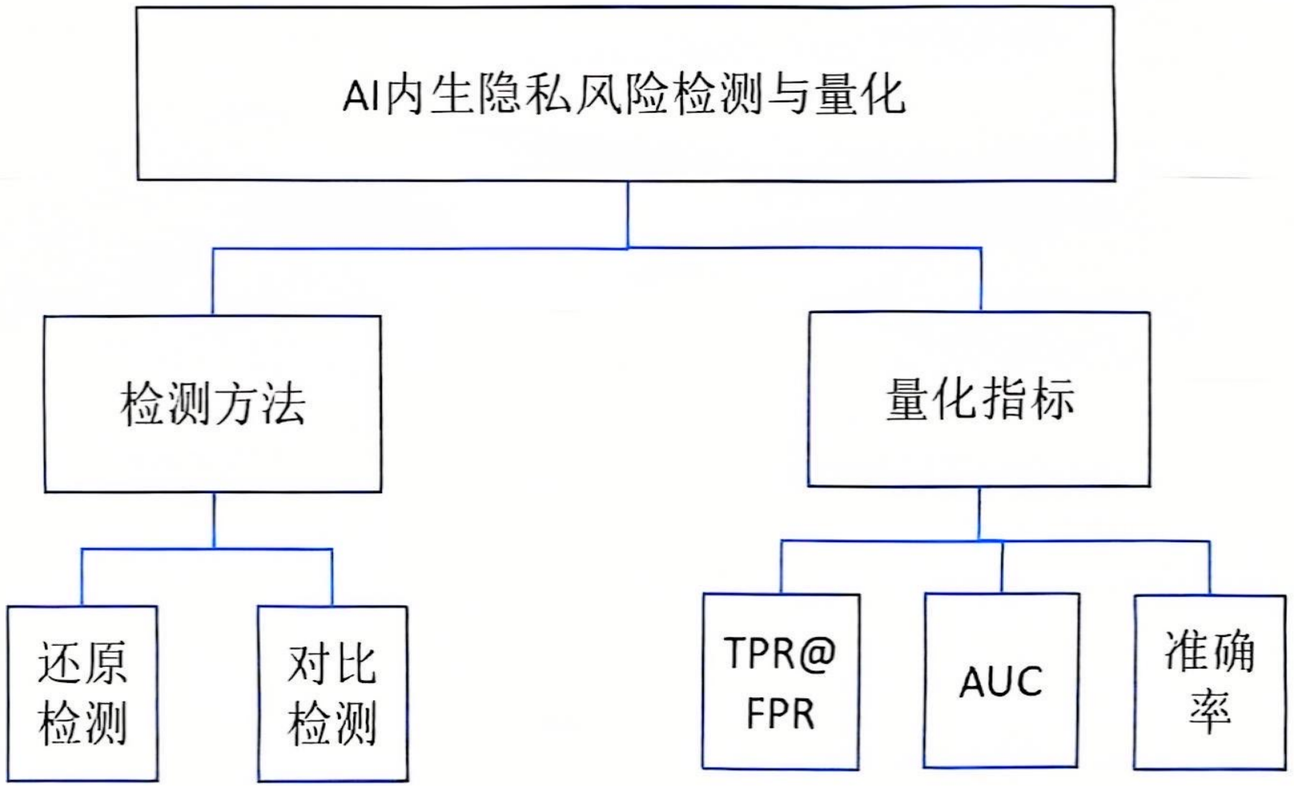

验证原型
验证反馈机制
AI内生隐私泄露风险评估
反馈
理论研究
AI内生隐私风险检测与量化
检测方法
量化指标
量化结果
理论支撑
AI内生隐私风险归因与组成
攻击原理
组成部分
定义范围
AI内生隐私风险定义与分类表征方法
敌手权限
攻击目标
隐私定义
反馈
AI治理和数据安全相关的法规标准和共识指引
隐私风险分级分类
数据
私人数据
公开数据集
。。。

同济大学
TONGJI UNIVERSITY

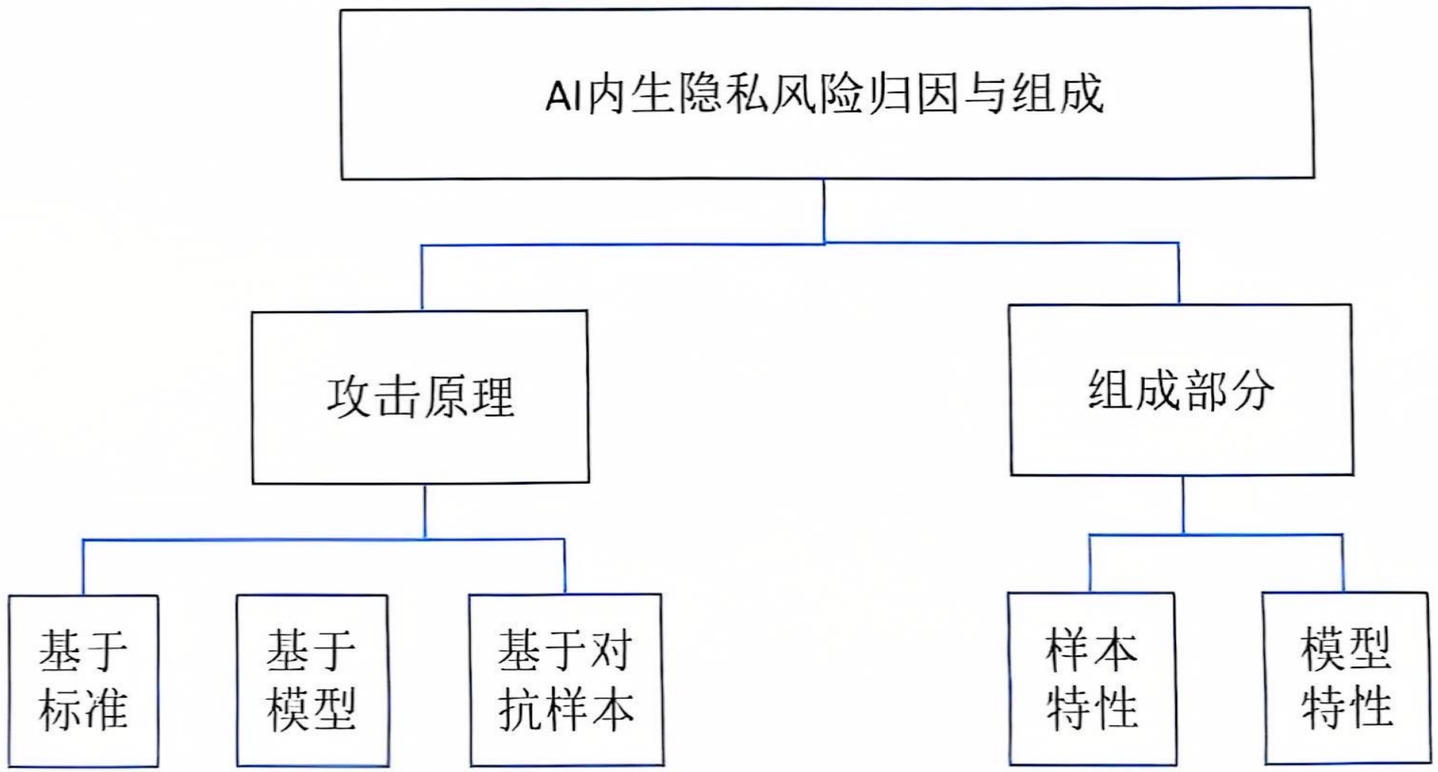
AI内生隐私风险定义与分类表征方法



AI内生隐私风险检测与量化



AI内生隐私风险归因与组成



团队研究进展

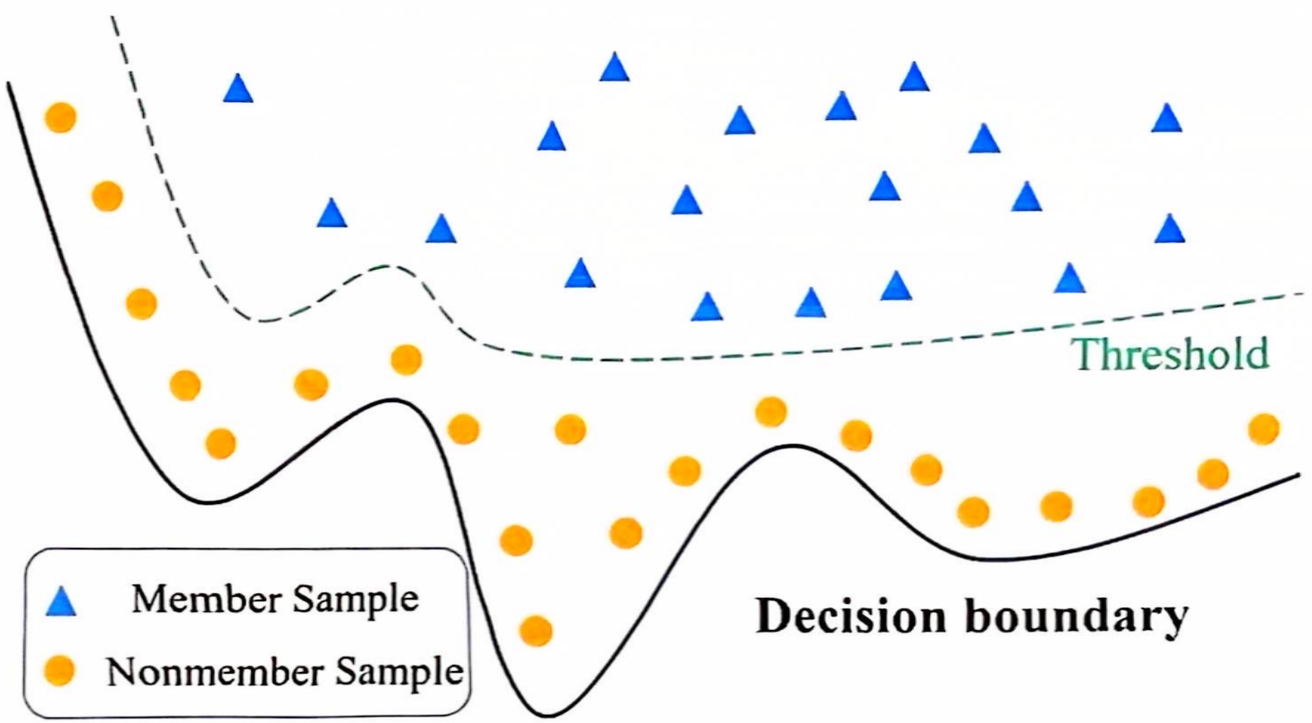
问题1	Optimizing label-only membership inference attacks by global relative decision boundary distances （ ISC 2024）
问题2	Unawareness detection: Discovering black-box malicious models and quantifying privacy leakage risks（Computers & Security）
问题3	Black-box membership Inference attack via knowledge distillation

成果一

传统纯标签成员推理攻击核心原理：成员相比于非成员更加难以被扰动，因此拥有更长的决策边界距离。

然而我们认为纯标签推理攻击实际隐私风险被低估了，主要由于以下两个问题：

- 现有攻击原理存在漏洞：成员与非成员的决策边界距离存在一定的重合部分
- 现有决策边界距离不稳定：获取决策边界距离方法随机性强，无法得到最优对抗点



成果一创新点

- 提出LDRSA算法，避免了由初始对抗点随机性带来的错误决策边界距离，得到了稳定的最优对抗点。

- 提出相对决策边界距离代替单一决策边界距离，改进攻击原理，增强了攻击的有效性。

实验结果

实验表明，我们的GAA攻击表现在2个数据集，三个指标上都优于其他攻击，表明我们的攻击效果造成了更大的隐私威胁。

此外，采用相对决策边界距离的攻击（UAA,TAA,GAA）表现都远优于采用决策边界距离的攻击（UA,TA,GA），这表明相对决策边界距离是一种更能区分成员与非成员区别的特征。

CIFAR10数据集			
攻击类型	TPR@0.1%FPR	TPR@1%FPR	AUC
UA	0%	1.4%	0.621
TA	0.2%	0.8%	0.621
GA	0.4%	0.8%	0.629
UAA	0.8%	2.6%	0.690
TAA	0.6%	2.8%	0.695
GAA	**1.4%**	**3.2%**	**0.703**

CIFAR100数据集			
攻击类型	TPR@0.1%FPR	TPR@1%FPR	AUC
UA	0.2%	0.6%	0.677
TA	0%	1.2%	0.678
GA	0%	0.6%	0.687
UAA	0%	0.2%	0.751
TAA	0.6%	1.6%	0.757
GAA	**1.6%**	**3.8%**	**0.765**

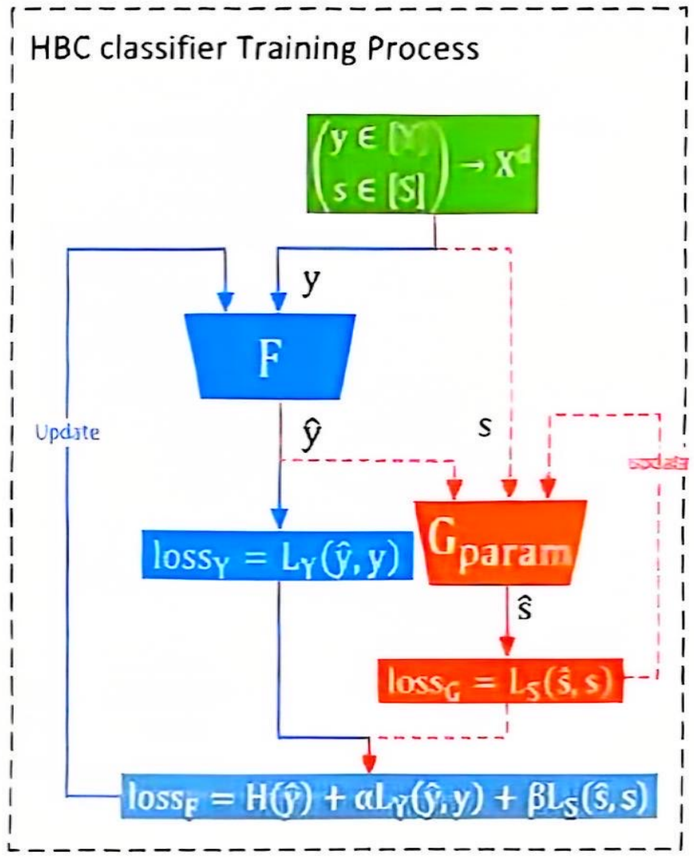
成果二

半诚实分类器（HBC）是一种典型的黑盒恶意模型，其目的为推理出训练样本中的敏感属性。

传统的还原检测方法难以应用于HBC模型，因为敌手与检测者的信息不对等：

- 传统评估模型输入和训练策略与HBC模型解码器存在差别，评估模型结果无法代表HBC模型真实造成的隐私泄露风险

- HBC模型隐私泄露风险受到数据集分布、模型结构、损失函数等多重因素影响，缺少对HBC分类器独有隐私风险的评估方法。

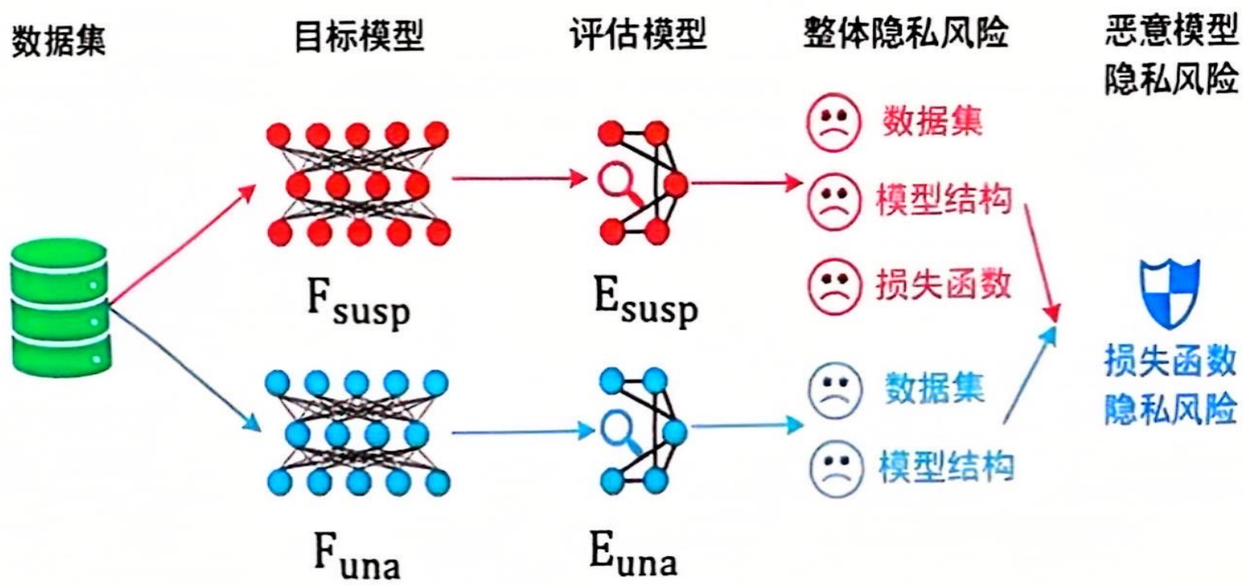


成果二

我们提出了一种基于UNA分类器的恶意模型检测法。

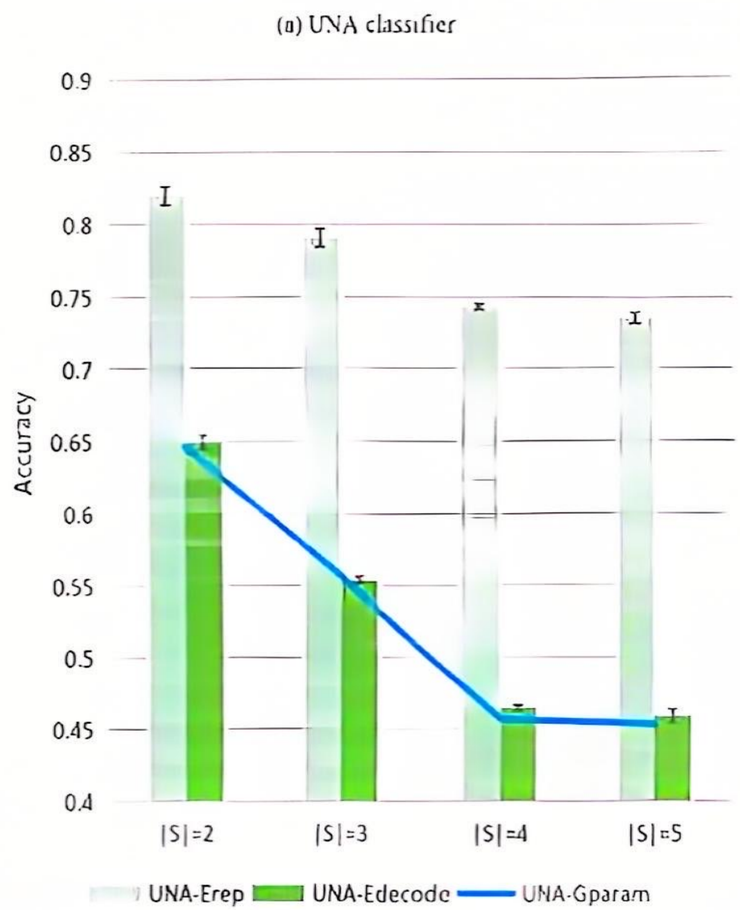
- 提出了新的评估模型，以目标模型最终输出为输入，更加贴合敌手解码器造成的实际敏感属性泄露

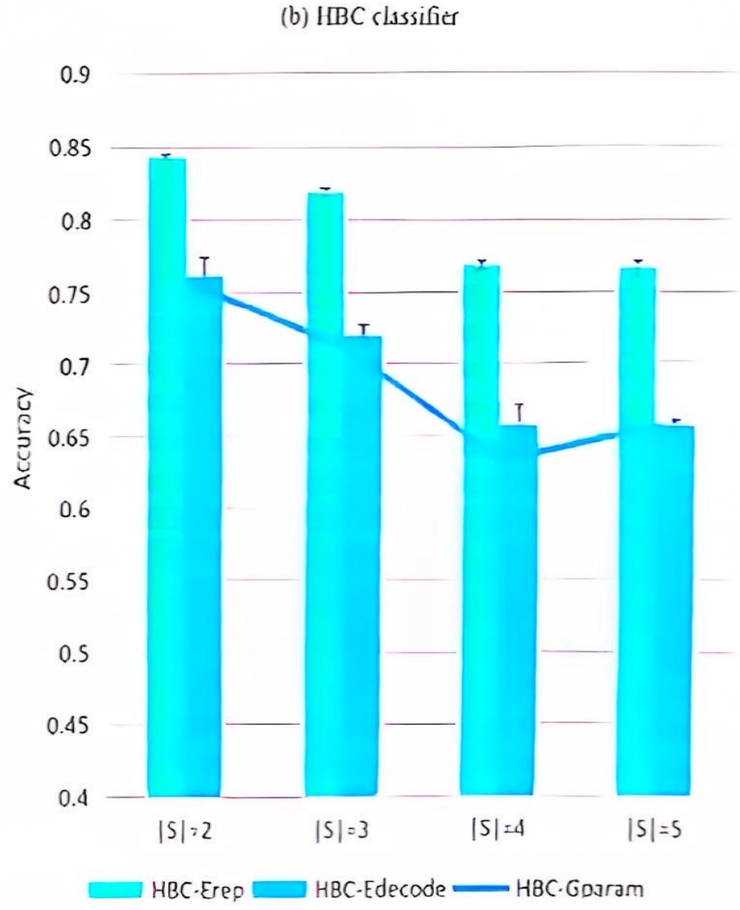
- 定义并提出损失函数不受敏感属性影响的UNA模型，对比后得到HBC分类器独有的隐私风险



实验结果

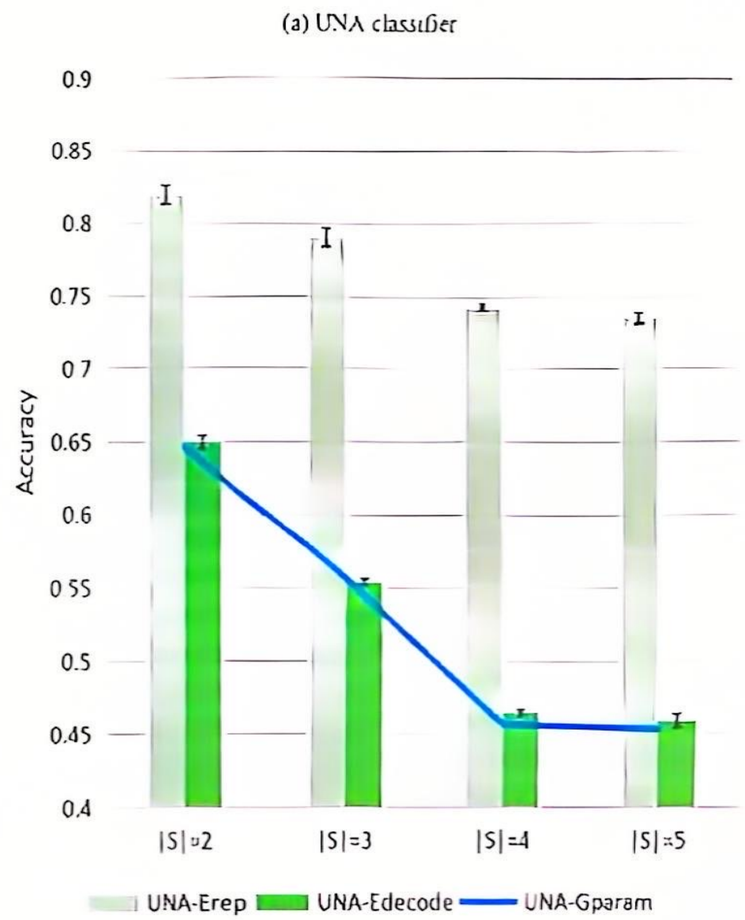
传统评估模型E_{rep}的准确率远高于解码器G_{param}准确率，而我们的评估模型E_{decode}的准确率与解码器G_{param}准确率（折线）十分接近，我们的评估模型更符合实际造成的隐私风险

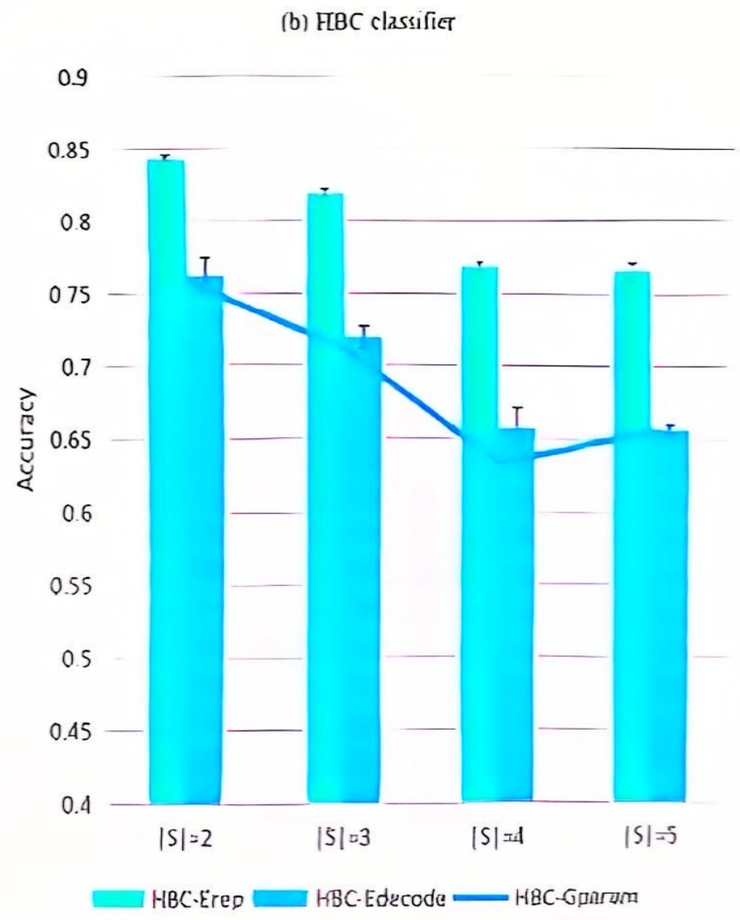




实验结果

传统评估模型E_{rep}的准确率远高于解码器G_{param}准确率，而我们的评估模型E_{decode}的准确率与解码器G_{param}准确率（折线）十分接近，我们的评估模型更符合实际造成的隐私风险



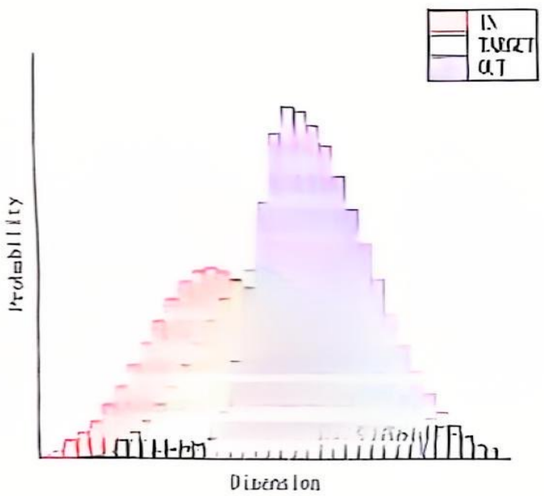


成果三

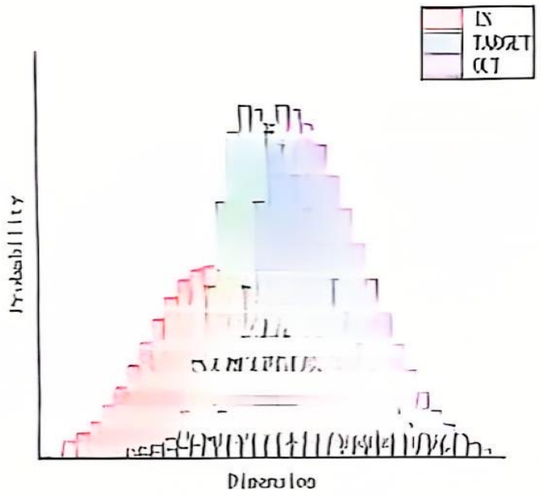
似然比攻击（LiRA）是目前最主流的成员推理攻击，其核心原理在于样本损失差距大的样本更容易泄露成员关系。

现有的LiRA主要存在两个缺陷：

- LiRA背后的直觉是经验性的，缺少对应的理论证明
- 传统LiRA训练需要大量影子模型，耗费海量时间和空间成本



成员样本



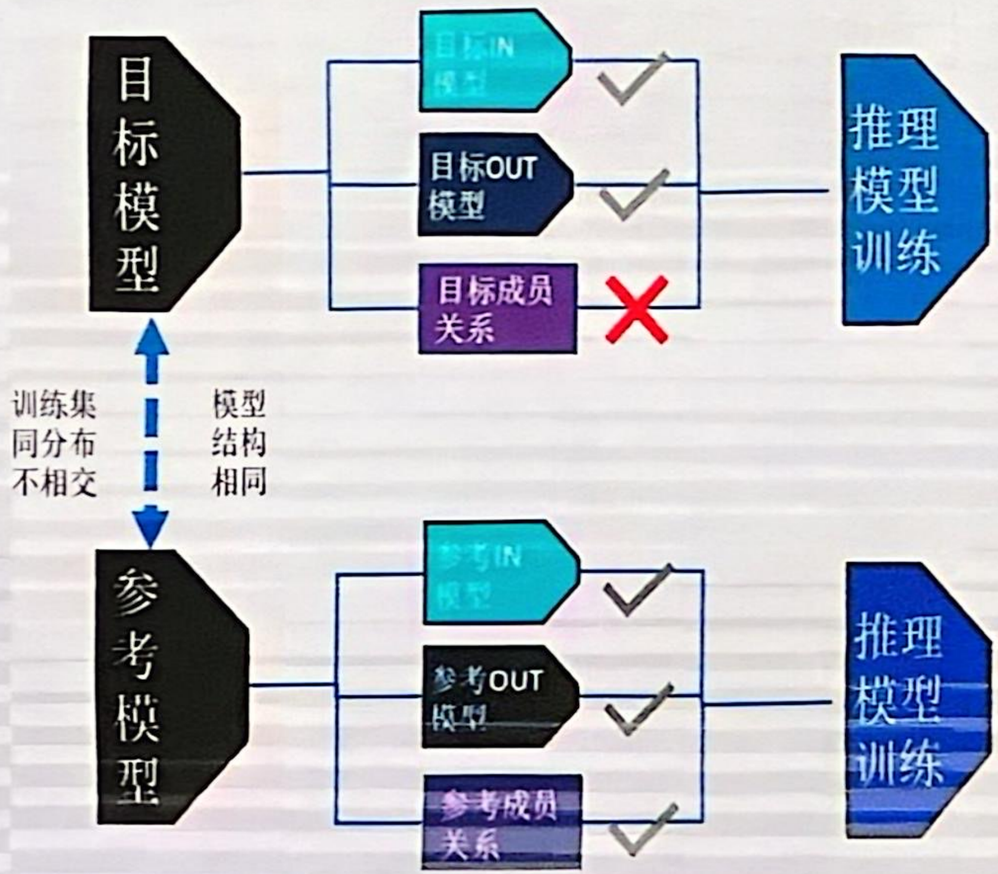
非成员样本

成果三

我们提出了一种基于知识蒸馏的推理攻击方法DIM-MIA。

- 提出并证明样本成员关系泄漏下限与样本损失差距之间存在正相关性

- 使用对目标模型知识蒸馏训练影子模型，使得影子模型输出更加符合样本参与目标模型训练前后的输出表现。



成果三创新点

- 提出DIM-MIA攻击方法，利用知识蒸馏来训练影子模型和参考模型，从而推理出成员关系，在有限的影子模型中实现了较高的攻击效率。

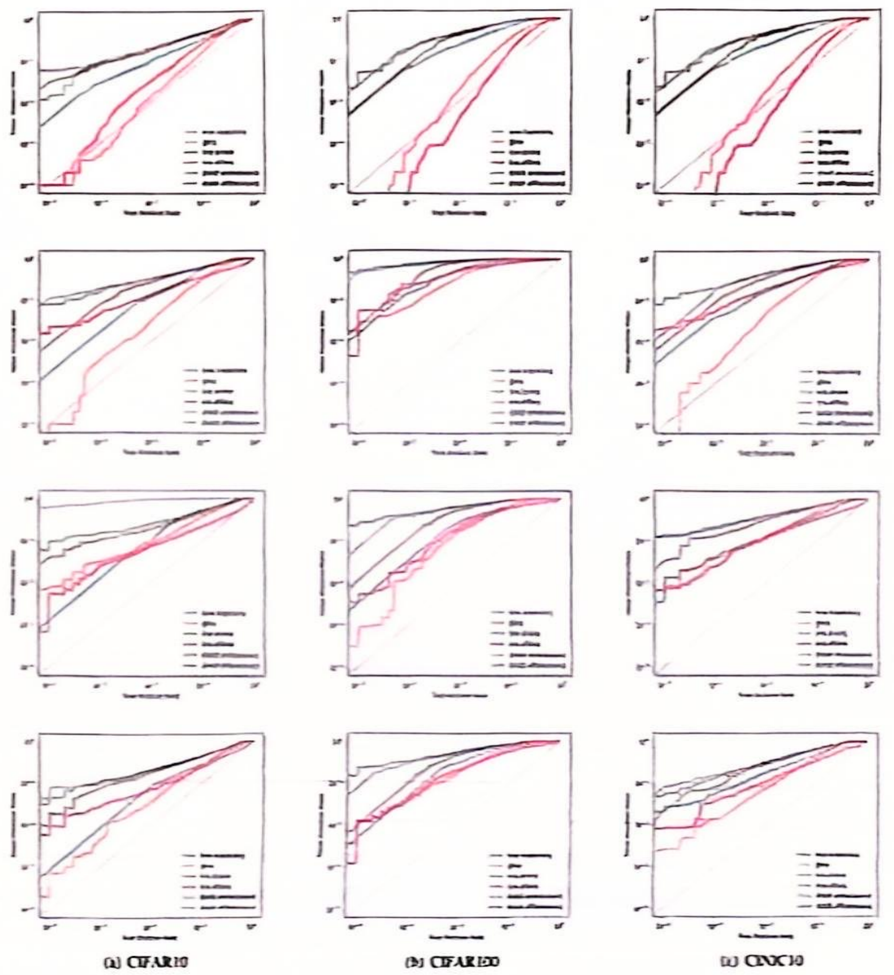
- 证明了样本成员泄漏下限与样本损失差距之间的正相关性，为过去的经验直觉提供了理论基础。

实验结果

在3种数据集4种模型结构下，我们限定影子模型数量为32进行实验。

我们的Distill online/offline攻击与LiRA online攻击远优于其他三种攻击。

Distill Offline攻击在不需要为每一个新来的样本额外训练IN模型，相比于传统的LiRA online攻击更具有普适性。



未来展望

当前人工智能模型在处理隐私泄露问题时，普遍将整个训练数据集视为单一的隐私保护对象。然而，训练数据集中各个样本的隐私敏感度存在差异，其隐私权重并不一致。

依据现行法律法规，对训练数据集数据进行精确的隐私权重分配，并将其反映在隐私风险评估结果中，是本领域未来研究的关键方向。