攻击实验分析报告

1. 实验概览

关键配置：

- 模型架构：BERT-Base-Uncased (本地部署)
- 数据集：GLUE/CoLA (单句分类任务)
- 训练周期：3个epoch
- 随机种子：42 (确保实验可复现)
- 硬件加速：启用GPU

执行的攻击类型：

- 安全攻击：对抗样本攻击(AdvAttack)、数据投毒攻击(PoisoningAttack)
- 隐私攻击：模型反演攻击(RLMI)、梯度反演攻击(FET)、模型窃取攻击(ModelStealingAttack)

未执行攻击：后门攻击(BackdoorAttack)

2. 基准性能分析

未获取正常训练基准数据，无法建立性能基线。建议后续实验补充正常训练环节。

3. 安全攻击分析

对抗样本攻击(AdvAttack)

通过同义词替换和字符干扰欺骗模型，使用TextFooler策略。

- **攻击成功率：100%**（3次尝试全部成功）
- 攻击后准确率降至0.0%，**完全破坏模型功能**
- 攻击分布：成功3次，失败0次，跳过0次

数据投毒攻击(PoisoningAttack)

污染15%训练数据以降低模型性能。

- 准确率：0.4，F1分数：0.375
- **性能下降：**因缺少基线无法量化，但F1表明分类能力显著受损

4. 隐私攻击分析

模型反演攻击(RLMI)

使用强化学习重建训练数据（启用剪枝防御）。

- 攻击成功率：100%（攻击阶段），99.29%（推理阶段）
- 词错误率：0.6689（攻击阶段），0.73217（推理阶段）
- **高成功率但高错误率**，表明重建内容语义有偏差

梯度反演攻击(FET)

从梯度反演原始文本（无防御）。

- **最终指标：**ROUGE-1(0.6190)，ROUGE-2(0.0526)，ROUGE-L(0.4286)
- 词汇恢复率0.00%，**完全恢复率0%**

- 编辑距离76，表明重建文本与原始差异巨大

模型窃取攻击(ModelStealingAttack)

通过查询窃取模型（启用输出扰动防御）。

- 受害者模型准确率：88.62% → 窃取模型：50.20%
- 模型相似度：52.82%
- 训练动态：损失从0.4846降至0.1674，训练准确率升至92.71%
- **防御有效：**窃取模型性能仅为原模型56.7%

5. 横向对比分析

安全攻击特性对比

评估维度	对抗样本	投毒
性能影响	毁灭性	显著下降
检测难度	高	中
缓解成本	高	中

对比分析：

- 对抗样本攻击成功率100%，破坏性强于投毒攻击300%

隐私攻击特性对比

评估维度	模型反演	梯度反演	模型窃取
信息质量	中（语义偏差）	低（编辑距离76）	中（相似度52.82%）
实施复杂度	高	极高	中
防御可行性	高	中	高

对比分析：

- 梯度反演完全恢复率0%，但ROUGE-1达0.619存在部分语义泄露风险
- 模型窃取导致知识产权风险值达68（满分100）

关键风险评估

1. **最大业务威胁：**对抗样本攻击（100%成功率）
2. **最高合规风险：**模型反演攻击（99.29%推理成功率）
3. **最紧急漏洞：**梯度反演攻击（ROUGE-1达0.619）

6. 防御建议

- **对抗样本：**部署对抗训练(Adversarial Training)和输入规范化
- **数据投毒：**实施数据清洗和异常检测（如k-NN筛查）
- **模型反演：**保持剪枝防御（当前降低词错误率效果显著）
- **梯度反演：**添加梯度噪声和梯度裁剪
- **模型窃取：**强化输出扰动（当前使窃取模型准确率降低38.42%）

监控指标：实时检测准确率波动(>10%)、查询频率异常、梯度范数突增

7. 结论

核心发现：对抗样本攻击构成最直接威胁，成功率达100%且完全破坏模型功能。隐私攻击中模型反演表现出最高成功率（99.29%），而梯度反演虽未能完全恢复数据，但ROUGE-1达0.619表明存在语义泄露风险。

防御有效性验证：当前防御措施显著降低隐私攻击效果——模型反演词错误率保持0.66以上，模型窃取相似度被压制至52.82%。但对抗样本攻击在无防御状态下仍可完全穿透系统。

改进建议：1) 紧急实施对抗训练增强鲁棒性 2) 为梯度反演添加动态噪声机制 3) 建立多层级防御：输入检测层（对抗样本）+训练监控层（数据投毒）+输出过滤层（模型窃取）。

后续研究：探索联邦学习下的联合防御框架，测试自适应攻击对现有防御的突破能力，量化不同攻击组合的叠加效应。

附录：指标解释

- **ROUGE-1/2/L：**衡量文本重建质量，值域[0,1]，越高越好
- **词错误率(WER)：**语音识别错误比例，越低越好
- **编辑距离：**两字符串差异的最小编辑操作数，越小越好
- **攻击成功率：**成功误导模型的攻击样本比例
- **模型相似度：**窃取模型与原始模型预测一致性