攻击实验分析报告

1. 实验概览

- **模型架构**: BERT-Base-Uncased (本地部署)
- **数据集**: GLUE/CoLA (单句分类任务)
- **训练参数**: 随机种子42, GPU加速, 正常训练3个epoch
- **执行的攻击类型**:
  - 安全攻击: 数据投毒攻击(PoisoningAttack)
  - 隐私攻击: 无有效数据

2. 基准性能分析

正常训练模型性能:

- 准确率(Accuracy): **40.0%**
- F1分数: **37.5%**

与基准对比: GLUE/CoLA任务中BERT-Base通常达到50-60%准确率(MCC指标)，本实验性能显著偏低，**表明模型未充分收敛或训练配置需优化**（仅3个epoch可能是主因）。

3. 安全攻击分析

数据投毒攻击(PoisoningAttack)

通过污染训练数据（投毒率15%）破坏模型性能，攻击参数：15个epoch，未启用防御。

- 攻击后准确率: 40.0% (与基准一致)
- 攻击后F1分数: 37.5% (与基准一致)
- **性能下降率: 0%** (未观测到有效攻击影响)

异常分析: 投毒攻击未造成性能退化，可能原因：

- 攻击样本设计缺陷，未触发模型决策边界变化
- 模型本身性能已接近随机水平(40%)，难以进一步降低
- 投毒策略与任务目标不匹配

4. 隐私攻击分析

未获取有效实验数据（RLMI/FET/ModelStealingAttack结果为空）

5. 横向对比分析

安全攻击特性对比

评估维度	投毒
性能影响	零影响（攻击失效）

评估维度	投毒
检测难度	高（未触发性能异常）
缓解成本	低（当前无需额外措施）

关键风险评估

1. **最大业务威胁**: 模型基础性能缺陷（非攻击导致）
2. **最高合规风险**: 数据完整性失效风险
3. **最紧急漏洞**: 训练流程缺陷（epoch不足）

6. 防御建议

- **投毒攻击防御**：当前攻击未生效，但建议：
  - 部署数据来源验证机制（如数据指纹）
  - 引入异常样本检测（如置信度过滤）
- **核心改进方向**：
  - 增加训练epoch至标准值（建议≥10）
  - 监控训练曲线确保收敛
  - 采用鲁棒优化器（如AdamW）

7. 结论

本次实验揭示了模型基础性能的严重缺陷：**正常训练准确率仅40%**，显著低于BERT在CoLA任务的典型表现（55-65%），表明训练配置（仅3个epoch）不足以支撑模型收敛。在投毒攻击测试中，**15%投毒率未对模型产生可观测影响**，这既可能反映攻击策略失效，更暴露模型本身性能已接近随机分类水平，缺乏进一步下降空间。

关键风险集中于**训练流程缺陷**而非外部攻击，建议优先：1) 延长训练epoch至10轮以上并监控收敛；2) 验证数据预处理流程；3) 引入学习率调度机制。后续实验需：首先优化基础性能至合理水平（准确率>55%），再重新评估攻击效果；同时建议补充对抗样本和后门攻击测试以全面评估安全态势。

未来研究应关注：1) 低性能模型的安全评估方法论；2) 投毒攻击在边缘分布样本中的有效性；3) 训练早期阶段的异常检测技术。

附录：指标解释

- **准确率(Accuracy)**: 正确预测样本占总样本的比例
- **F1分数**: 精确率与召回率的调和平均数，综合反映分类性能
- **投毒率(Poisoning Rate)**: 污染样本占训练集的比例