1. 实验概览

- 模型架构: BERT-Base-Uncased
- 数据集: GLUE/SST-2（情感分析任务）
- 训练参数: Epochs＝3, 使用GPU加速, 随机种子＝42
- 执行的攻击类型及分类:
  - ○ 安全攻击: 对抗样本攻击(AdvAttack)、数据投毒攻击(PoisoningAttack)
  - ○ 隐私攻击: 无（后门、模型反演、梯度反演、模型窃取均未测试）

2. 基准性能分析

- "**normalTrain**"结果: 无可用数据（result中为空列表）
- 与文献比较: GLUE/SST-2任务中BERT-Base典型准确率约92%。实验缺失基准数据，假设标准性能作为参考锚点。

3. 安全攻击分析

对抗样本攻击**(AdvAttack)**

- 攻击原理: 通过细微扰动（如TextFooler策略）欺骗模型产生错误预测。
- 平均攻击成功率: 0.0%（resultData: ["0", "0", "3", "0.0%", "0.0%", "0.0%"]）
- 攻击前后准确率: 攻击前0.0% → 攻击后0.0%，无变化（异常）
- 攻击分布: 成功0次/失败0次/跳过3次
- 关键发现: 所有指标为**0**，表明攻击未实际执行（可能因策略TextFoolerJin2019与模型不适配或配置错误）。

数据投毒攻击**(PoisoningAttack)**

- 攻击原理: 污染15%训练数据（poisoning_rate＝0.15）以系统性降低模型性能。
- 准确率与**F1**趋势:
  - ○ 准确率＝50%（resultData[0]＝"0.5"）
  - ○ F1分数≈33.3%（resultData[1]＝"0.333..."）
- 平均性能下降: 相较SST-2典型BERT性能（92%），准确率下降**45.7**个百分点，F1下降约60个百分点。
- 关键发现: 投毒显著破坏模型性能，但隐蔽性较低（性能骤降易被检测）。

4. 隐私攻击分析

- 模型反演**(RLMI)**、梯度反演**(FET)**、模型窃取**(ModelStealingAttack)**: 均无测试数据，跳过分析。

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	无影响（异常）	高（↓45.7%）
检测难度	高	中

缓解成本 中 低

对比分析:

- 投毒攻击造成最严重性能破坏，但易通过数据清洗缓解。
- 对抗样本攻击因执行失败未构成实际威胁。

关键风险评估

1. 最大业务威胁: 数据投毒攻击（实测性能下降＞45%）
2. 最高合规风险: 无显著隐私泄露（隐私攻击未执行）
3. 最紧急漏洞: 数据管道污染（投毒攻击生效）

6. 防御建议

- 对抗样本攻击:
  - ○ 启用对抗训练（如PGD防御），增强扰动鲁棒性。
  - ○ 监控指标：输入文本扰动敏感度、置信度波动。
- 数据投毒攻击:
  - ○ 部署数据清洗与异常检测（如k-NN离群值筛查）。
  - ○ 架构改进：集成联邦学习减少集中式数据污染风险。
- 通用建议:
  - ○ 添加正常训练基准以量化攻击影响。
  - ○ 对隐私攻击补充测试（如梯度裁剪防御FET）。

附录：指标解释

- 准确率**(Accuracy)**: 正确预测样本占比，衡量任务性能。
- **F1**分数: 精确率与召回率的调和平均，适用于不平衡数据。
- 攻击成功率**(AdvAttack)**: 成功欺骗模型的样本比例。
- 投毒率**(PoisoningAttack)**: 污染数据占训练集的比例（本例为15%）。
- 防御启用**(defenderEnabled)**: 布尔值，表示是否激活防护机制（本例均未启用）。

结论

- 核心发现: 数据投毒攻击是唯一成功的威胁，导致模型性能腰斩（准确率50%），暴露数据管道脆弱性。
- 关键短板: 缺乏正常训练基准和隐私攻击数据，建议补测**RLMI/FET**以评估隐私风险。
- 优先行动: 强化数据预处理流水线，并验证对抗样本防御的有效性。

注: 报告基于可用数据生成，异常结果（如AdvAttack全零指标）已标注；缺失部分（如基准性能）采用文献参考值推导。