1. 实验概览

- **模型架构**：BERT-Base-Uncased
- **数据集**：IMDB（电影评论情感分类）
- **训练参数**：本地数据集训练，未启用GPU，随机种子42
- **执行攻击类型**：
  - **安全攻击**：对抗样本攻击（AdvAttack）、后门攻击（BackdoorAttack）、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）

2. 基准性能分析

- 正常训练准确率：**90.0%，F1分数：0.859**
- 对比文献基准（BERT-Base在IMDB任务）：略低于典型基准（通常>92%），**表明模型初始性能存在优化空间**

3. 安全攻击分析

- **对抗样本攻击（AdvAttack）**：
  通过同义词替换干扰文本欺骗模型。
  - 平均攻击成功率：**65.0%**（两次攻击一致）
  - 攻击后准确率暴跌至35.0%（防御启用时攻击前100.0%→35.0%，未防御时98.0%→35.0%）
  - 攻击分布：成功25次/失败15次（无跳过）
  - **关键发现**：防御机制对攻击成功率无显著影响，**暴露模型鲁棒性缺陷**
- **后门攻击（BackdoorAttack）**：
  植入触发词操纵模型输出。
  - 毒化后准确率骤降（0.883%→0.277%）
  - 隐蔽性指标：困惑度/语义相似性/语法正确性均为 `NaN`（**数据异常，无法评估隐蔽性**）
  - 毒化效果雷达图（概念）：

```
radarChart
    title 毒化效果评估
    axis 准确率，困惑度，语义相似性，语法正确性
    "原始模型" [0.883, 0, 0, 0]
    "毒化模型" [0.277, NaN, NaN, NaN]
```

- **数据投毒攻击（PoisoningAttack）**：
  污染训练数据降低模型性能。
  - 攻击后准确率：**50.0%（基准90.0%），F1分数：0.759**
  - 性能下降幅度：**44.4%**（相对基准）
  - **异常点**：攻击配置中 `executed: false`，但结果数据存在，**需验证执行真实性**

4. 隐私攻击分析

- **模型反演攻击（RLMI）**：
  重构训练数据特征。
  - 攻击/推理阶段成功率均达**100.0%**
  - 词错误率：攻击阶段0.794→推理阶段0.720，**显示攻击过程优化效果**
  - **矛盾点**：高成功率与高词错误率并存，**暗示重构数据质量不稳定**
- **梯度反演攻击（FET）**：
  从梯度泄露原始数据。
  - 最终指标：ROUGE-1=0.728，ROUGE-L=0.437，词汇恢复率=0.0，编辑距离=85.0
  - 训练动态：ROUGE-L从0.437（Epoch1）无后续数据
  - **关键异常**：词汇恢复率=0.0但编辑距离高达85，**表明重构文本表面相似但语义失真**
  - **转折点**：仅单Epoch数据，无法分析趋势
- **模型窃取攻击（ModelStealingAttack）**：
  复制受害者模型功能。
  - 性能对比：受害者模型准确率88.62%→窃取模型50.20%
  - 模型相似度：**52.84%**（较低）
  - 训练过程：损失从0.48降至0.15，训练准确率从86.25%升至92.71%
  - **防御影响**：启用防御时受害者模型性能更高（88.62% vs 未防御80.0%）

5. 横向对比分析

安全攻击特性对比

评估维度	对抗样本	后门	投毒
性能影响	高（-65% ACC）	极高（-99.7%）	高（-44.4%）
检测难度	中	低（异常显著）	中
缓解成本	高	低	中

对比分析：

- 后门攻击破坏强度**超行业平均阈值（>95%性能下降）**
- 对抗样本攻击成功率**65%**，需优先防御

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	低（高错误率）	极低（语义失真）	中（功能部分复制）
实施复杂度	高	极高	中
防御可行性	中	低	高

对比分析：

- 模型窃取导致**知识产权风险值52.84%**，构成最高合规风险
- 梯度反演完全恢复率**0%**，但存在梯度泄露隐患

关键风险评估

1. **最大业务威胁**：对抗样本攻击（高频高破坏）
2. **最高合规风险**：模型窃取（知识产权侵害）
3. **最紧急漏洞**：后门攻击（近乎完全瘫痪模型）

6. 防御建议

- **对抗样本**：
  - 部署对抗训练（Adversarial Training）
  - 监控指标：**输入文本扰动敏感度**
- **后门攻击**：
  - 采用异常激活检测（如Neuron Cleanse）
  - 架构改进：**输入过滤层拦截触发词**
- **模型窃取**：
  - 限制API查询频率
  - 监控指标：**相同查询的响应一致性**
- **通用建议**：
  - 梯度裁剪抑制梯度反演
  - 差分隐私增强数据安全

> *附录：指标解释*
>
> - ***ROUGE-L**：衡量生成文本与参考文本的最长公共子序列相似度*
> - ***词错误率（WER）**：语音/文本重构错误比例*
> - ***模型相似度（agreement）**：窃取模型与受害者模型输出一致性*
> - ***PPL（困惑度）**：语言模型预测能力指标，值越低越好*
> - ***攻击成功率**：对抗样本成功误导模型的比例*