1. □验概览

- 关键配置提取□
  - 模型架构□BERT-Base-Uncased
  - 数据集□GLUE\cola（CoLA 语料库，用于单句分类任务）
  - 训练参数□训练轮数（epochs）= 3，随机种子（random_seed）= 42，启用 GPU 加速（use_gpu=true）
  - 任务类型□单句分类（TaskForSingleSentenceClassification）
- 执行的攻击类型及分类□
  - 安全攻击□数据投毒攻击（PoisoningAttack）
  - 隐私攻击□无（模型反演、梯度反演、模型窃取攻击均未执行）

2. 基准性能分析

- **normalTrain** 结果分析□

  实验数据中未包含正常训练（normalTrain）的结果（result 为空列表）。

  无法提供基准准确率和 **F1** 分数，导致后续攻击影响评估受限。
- 与文献基准比较□

  在 GLUE\cola 数据集上，BERT-Base 模型的典型基准准确率约为 0.6–0.7（文献参考□Devlin et al., 2019）。

  由于缺少本地基准数据，无法直接对比，但投毒攻击结果显著低于此范围，暗示潜在攻击成功。

3. 安全攻击分析

- **数据投毒攻击(PoisoningAttack)**□
  - 攻击解释□通过污染训练数据（如注入恶意样本）降低模型性能，本次攻击投毒率为 15%（污染 15% 训练数据）。
  - 结果分析□
    - 攻击执行一次，准确率 = 0.4，F1 分数 = 0.375。
    - 性能变化趋势□无多次攻击数据，但当前值远低于文献基准（0.6–0.7），表明 攻击导致模型性能严重下降。
    - 平均性能下降□因缺少基准数据，无法量化，但绝对低值（<0.5）显示攻击有效。
  - 异常标注□准确率与 F1 分数均异常低（<0.5），可能因高投毒率（15%）或任务敏感度导致。

4. 隐私攻击分析

- 模型反演攻击**(RLMI)**、梯度反演攻击**(FET)**、模型窃取攻击**(ModelStealingAttack)**:

  □验数据中均未执行（**result** 为空列表），跳过分析。

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	显著下降（准确率 0.4）
检测难度	中等（隐蔽性较高）
缓解成本	中等（需数据清洗）

对比分析□

- 投毒攻击隐蔽性指数较高（因数据污染不易实时检测），但性能破坏强度突出（准确率仅 0.4）。

隐私攻击特性对比

无可用数据，省略表格。

关键风险评估

1. 最大业务威胁□模型性能崩溃（准确率降至 0.4，任务失效风险高）

2. [illegible]
3. [illegible]

6. [illegible]

- [illegible]
  - [illegible]
  - [illegible]
  - [illegible] 0.4[illegible]
- [illegible]
  - [illegible]
  - [illegible]
- [illegible]
  [illegible]

> [illegible]
>
> - **[illegible] (Accuracy)**[illegible] ([illegible] / [illegible])[illegible]
> - **F1 [illegible] (F1 Score)**[illegible] (2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}})[illegible]
> - **[illegible] (Poisoning Rate)**[illegible] 15%[illegible]