情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

AI生成したプログラムコードの識別に関する研究

大久保研究室
5541106 西村太孝
修士論文審査会

アジェンダ

- 研究背景
- 研究目的
- 既存研究とその限界
- 提案手法
- 実験結果
- 考察
- まとめと今後の課題

アジェンダ

- 研究背景
- 研究目的
- 既存研究とその限界
- 提案手法
- 実験結果
- 考察
- まとめと今後の課題

研究背景

- 生成AIによるコード生成の普及
- シャドーAI問題
- 蒸留モデルによる生成特性の類似化

- AI生成かどうか→不十分
- 必要なのは
  - どのAIが生成したか
  - 生成AI同士はどれくらい似ているか

アジェンダ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 研究背景
- **研究目的**
- 既存研究とその限界
- 提案手法
- 実験結果
- 考察
- まとめと今後の課題

研究目的

■目的

- 生成元識別と類似性の分析

■研究課題

① RQ1:生成AIモデルをコードから識別できるか、表現方法の影響はあるのか

② RQ2:生成AIモデル間の類似性は埋め込み空間上にどのように現れるか

貢献

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

■本研究の貢献

- Code-T5+110M を用いたコード埋め込みに基づき，生成AIコードの分類と類似性評価を同一の埋め込み空間上で総合的に扱う分析枠組みを提案した.
- 正規化処理，AST表現，およびUniXcoder を用いた構造表現を導入し，それらが生 成AIコード分析に与える影響を体系的に評価した.
- 異なる系列の生成AIモデルや，蒸留関係にある生成AIモデルを含む複数モデルを 対象として，モデル間の関係性がコード埋め込み空間上に現れることを実証的に示 した.

アジェンダ

- 研究背景
- 研究目的
- **既存研究とその限界**
- 提案手法
- 実験結果
- 考察
- まとめと今後の課題

既存研究とその限界

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

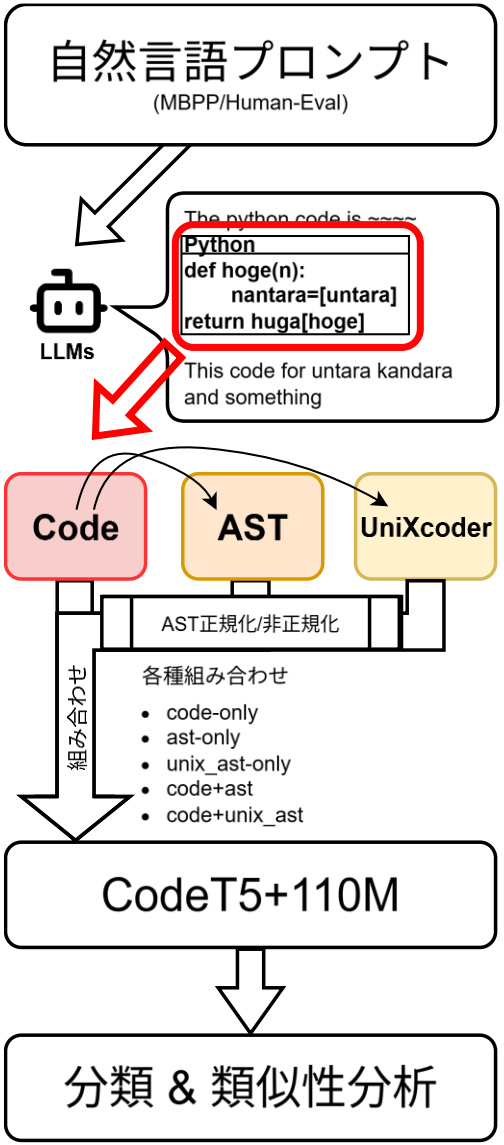
- Suhらの研究では
  - 自然言語AIGC検出器
  - コード特化検出器
  - CpdeT5/機械学習分類器
  - 限界
    - AI・ヒトの二分類が中心
  - この実験で一番精度が良かったのは CodeT5+埋め込みと機械学習分類器の組み合わせ

アジェンダ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 研究背景
- 研究目的
- 既存研究とその限界
- **提案手法**
- 実験結果
- 考察
- まとめと今後の課題

提案手法



1. 同一プロンプトでコード生成
2. Code-Only抽出
3. AST/UniXcoder ASTの適用
4. AST正規化有無の組み合わせ
5. CodeT5+を用いた埋め込み
6. 分類・類似性分析

提案手法

① コード表層情報
- 変数名や書式など、生成モデル固有の表層的書き癖を捉える

② 構文情報
- 表層表現に依存しない構文情報の差異を確認

③ AST表現方式
- AST表現方式の違いが識別性能に与える影響を検証

④ 正規化処理の有無
- 正規化処理によって失われた表層情報がどのような影響を与えるか

コード表現の設計

- Code-Only
- AST-Only
- Unix AST-Only（UniXcoder-AST）
- Code+AST
- Code+Unix AST

  ＋AST正規化の有無

AST正規化


```
def bell_number(n):
    # (n+1) x (n+1) の2次元配列を初期化
    bell = [[0 for _ in range(n + 1)] for _ in range(n + 1)]
    bell[0][0] = 1

    for i in range(1, n + 1):
        # 前の行の最後の値を、新しい行の先頭にコピー
        bell[i][0] = bell[i - 1][i - 1]

        for j in range(1, i + 1):
            # 左上と左の値を足し合わせる
            bell[i][j] = bell[i - 1][j - 1] + bell[i][j - 1]

    return bell[n][0]

def func(arg0):
    var0 = [[0 for var1 in range(arg0 + 1)] for var1 in range(arg0 + 1)]
    var0[0][0] = 1

    for var2 in range(1, arg0 + 1):
        var0[var2][0] = var0[var2 - 1][var2 - 1]

        for var3 in range(1, var2 + 1):
            var0[var2][var3] = var0[var2 - 1][var3 - 1] + var0[var2][var3 - 1]

    return var0[arg0][0]
```

AST正規化

変数名や関数名、書式を一意な連番に変更

コード自体の構造は変えない

実験設定

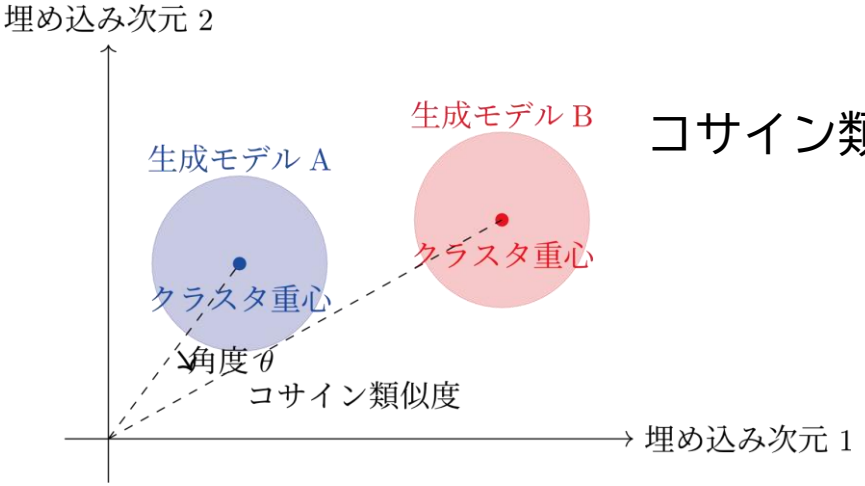
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- データ（プロンプト）
  - MBPP＋HumanEval-X（合計1096問）
- Temprature: 0/Default
- モデル
  - GPT-4o-mini
  - Deepseek-r1_32b,Qwen2.5_32b
  - Llama3,3.1,3.3
  - Gemini 2.0flash,2.5flash,2.5pro

・ 赤枠でかこまれたモデル（3種類）はRQ1の分類タスクで利用したモデル

実験設定

- データ分割
  - RQ1では、データ分割を80学習20評価
  - ランダム分割で10回繰り返し
- 類似度
  - コサイン類似度を用いて分析



コサイン類似度のイメージ

実験設定

- コード埋め込みの作成
  - CodeT5+
- 使用した分類器
  - Suhらの研究と同じく、
    LR,KNN,SVM,MLP,RF,DT,GB,XGBの８種類を使用
  - RQ1の結果として出てくる数値はこれらの平均

アジェンダ

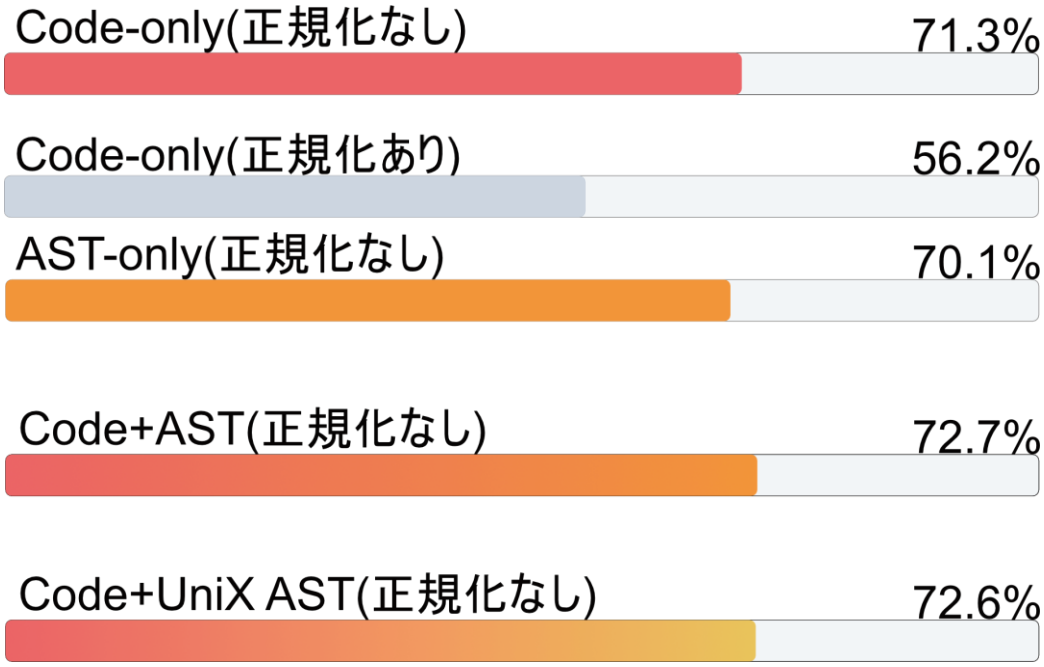
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 研究背景
- 研究目的
- 既存研究とその限界
- 提案手法
- **実験結果**
- 考察
- まとめと今後の課題

実験結果①（RQ1）

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- AST系表現の組み合わせが有効
- 正規化によって精度が低下
  - 削除した変数名、関数名、書式が割と大事



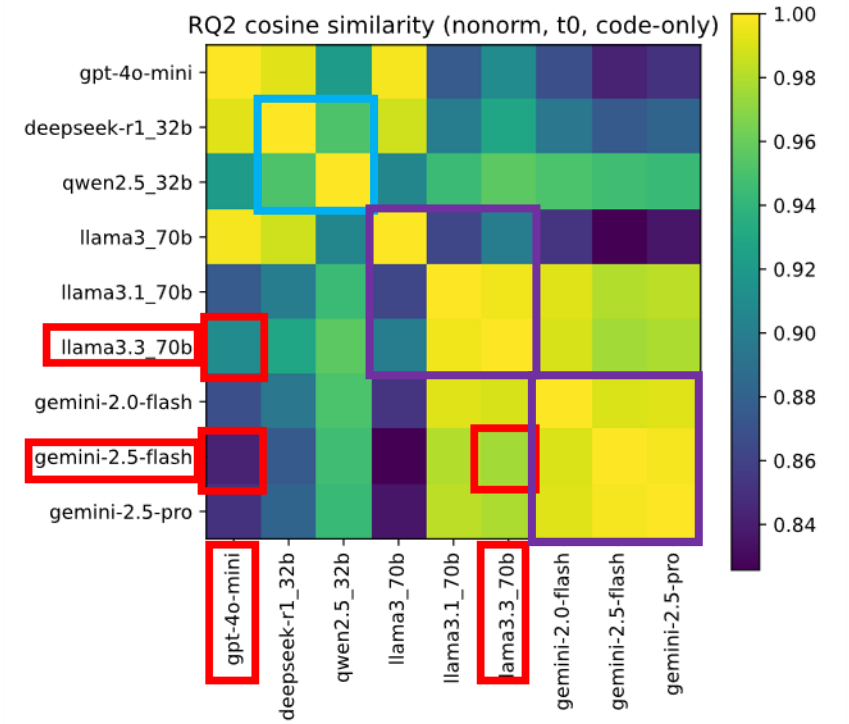
実験結果①（RQ1）

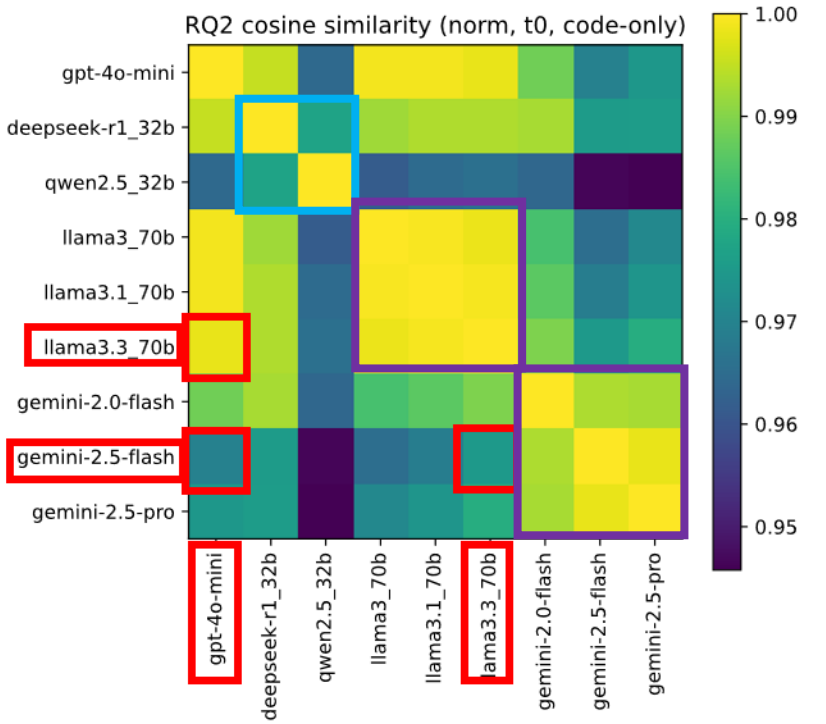
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 全体的な分類性能は低下
- 性能関係はT=0の場合と概ね一致
  - 生成時のゆらぎにもある程度耐えられる

	t0	td
Code	0.713	0.700
Code(正)	0.562	0.599
AST	0.701	0.693
Code+AST	0.727	0.714
Code+Unix	0.726	0.710

実験結果②（RQ2）





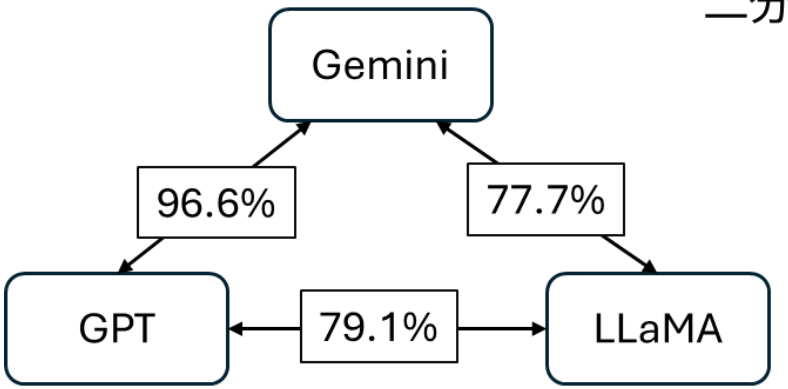
- 系列モデルはクラスタ化
- 蒸留モデルは近く（0.951→0.987）
- ただし、差は非常に小さい（最大0.03程度）

AST正規化の影響（RQ2）

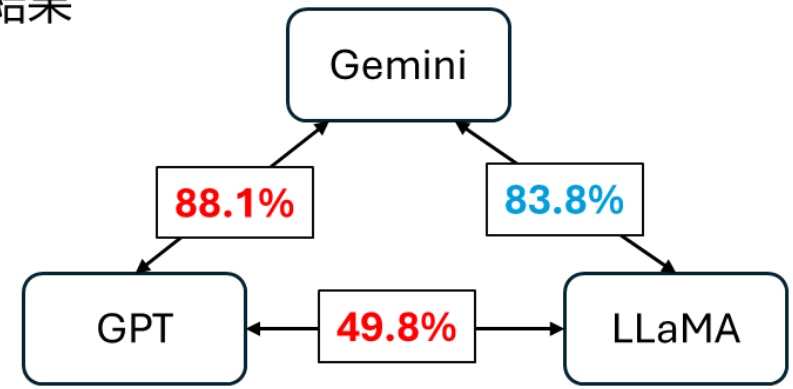
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 前ページ赤枠はRQ1で使用したモデル
- 埋め込み空間上での近接/遠隔が確認

二分類の結果



正規化ナシ



正規化アリ

埋め込み空間上で近接した分類は精度が下がり
遠隔した分類は精度が上がっている。

アジェンダ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 研究背景
- 研究目的
- 既存研究とその限界
- 提案手法
- 実験結果
- **考察**
- まとめと今後の課題

考察（RQ1）

- Code-OnlyよりもAST系列を組み合わせるといい
  - 表層的な情報と構造的情報の組み合わせ
- 正規化すると精度が落ちる
  - 分類では表層情報が大事
- Temperatureの影響
  - T=0の方が分類性能◯

考察（RQ2）

- モデル系列ごとの類似構造
  - 同系列は異系列と比べて類似度アップ
- 蒸留モデルの位置づけ
  - 不十分なところが多いのでさらなる調査が必要

アジェンダ

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 研究背景
- 研究目的
- 既存研究とその限界
- 提案手法
- 実験結果
- 考察
- まとめと今後の課題

まとめと今後の課題

- 表層情報は分類で有効
- モデル関係性をある程度分析可能
- 新しい分析枠組みを提案

- 他言語での適用
- 精度の向上