

Final project

Paper R&D Practice

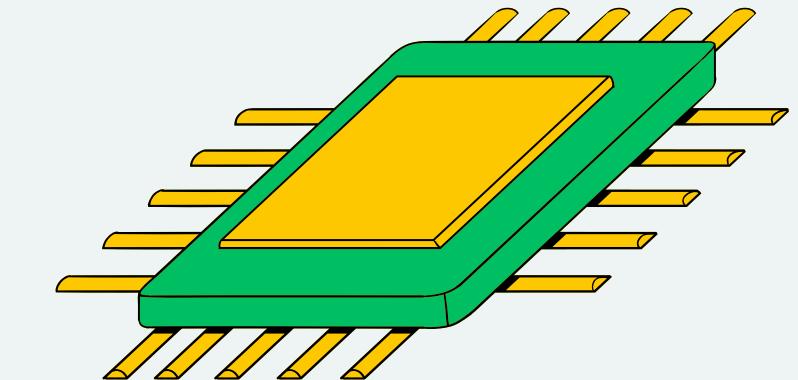
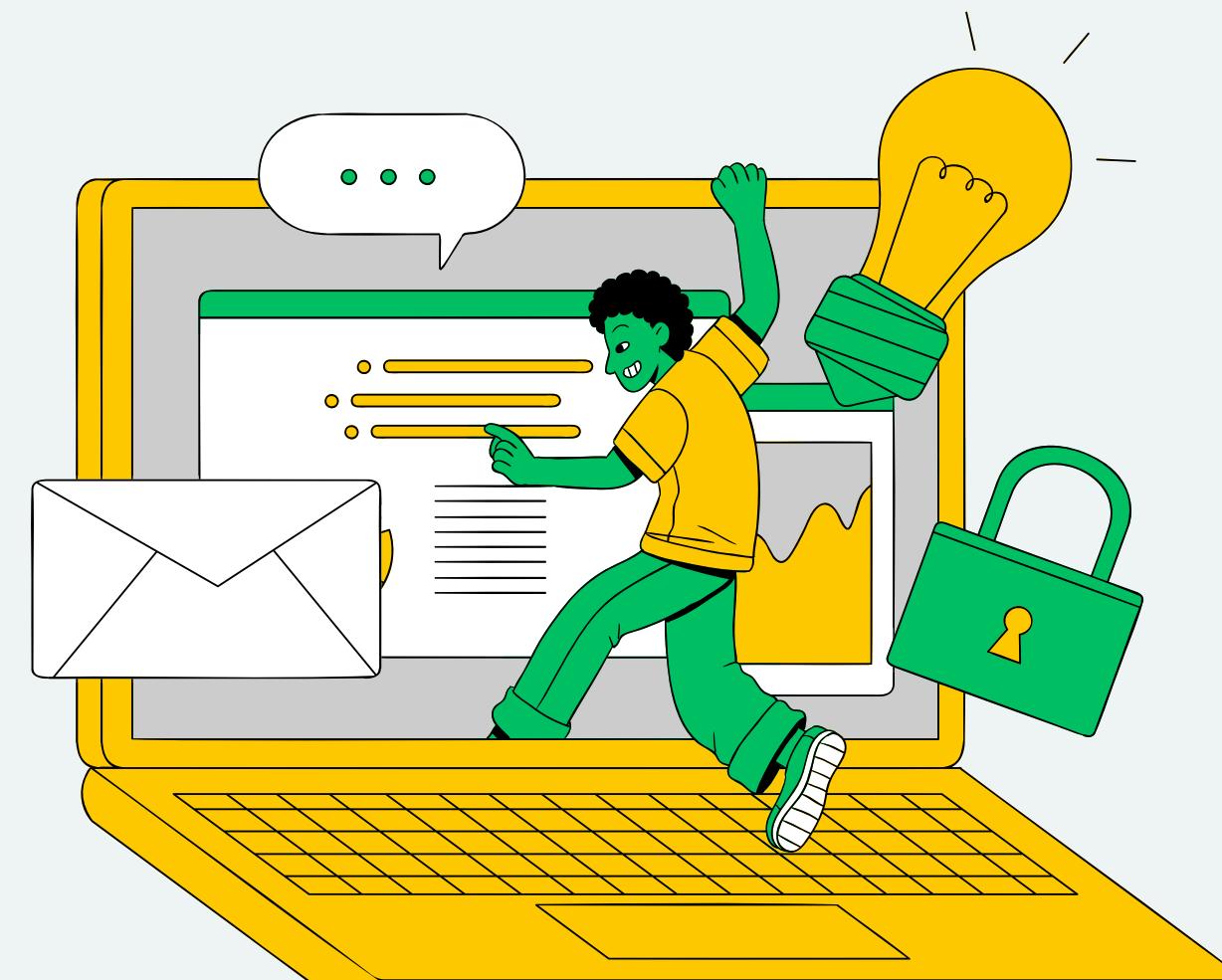
Paper: "Evaluating spam filters and Stylometric Detection of AI-generated phishing emails"

PRESENTED BY:

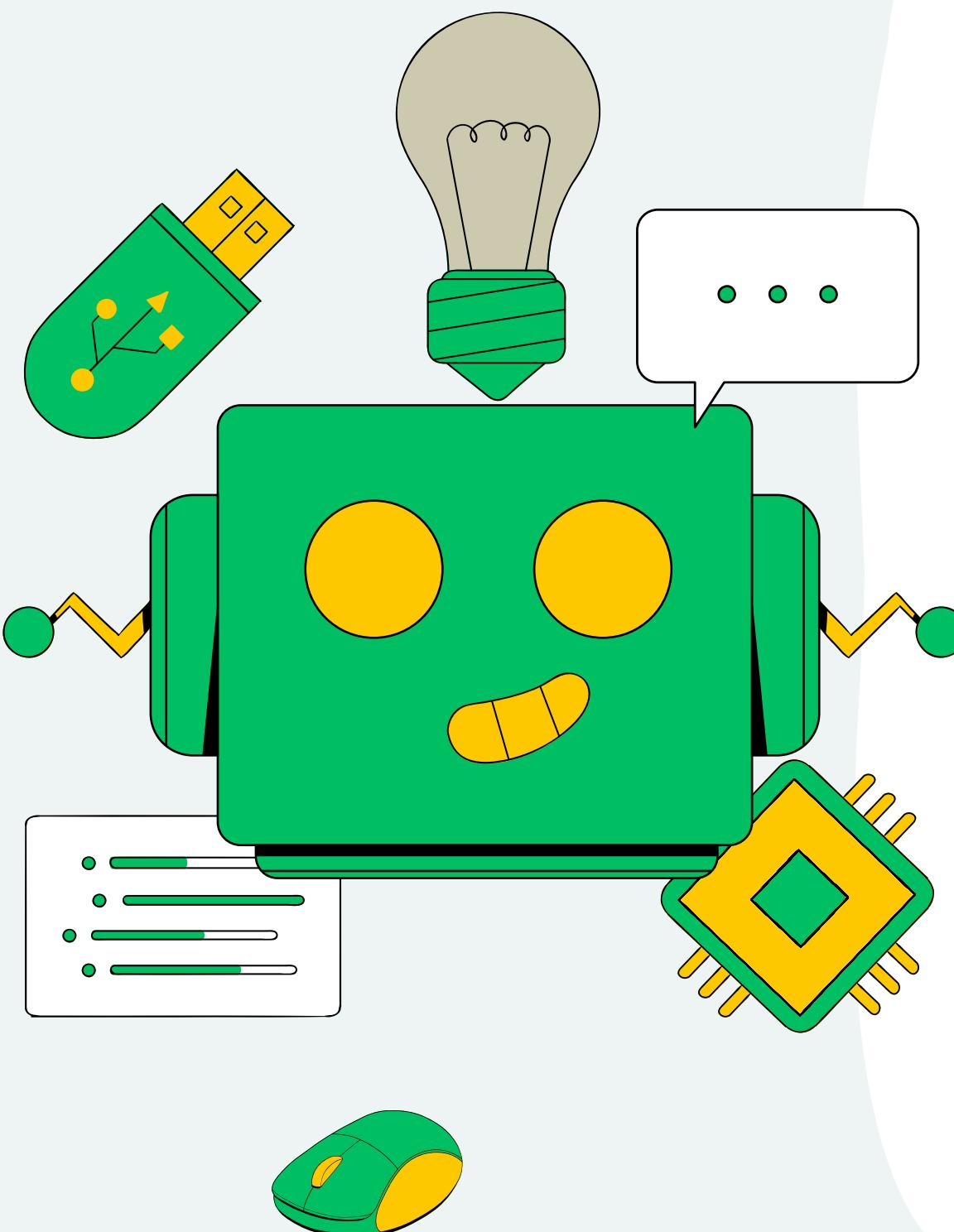
ณัฐริวิกาดา จิราธน์ชัย 6720422004

จรายพร ไหลกุล 6720422022

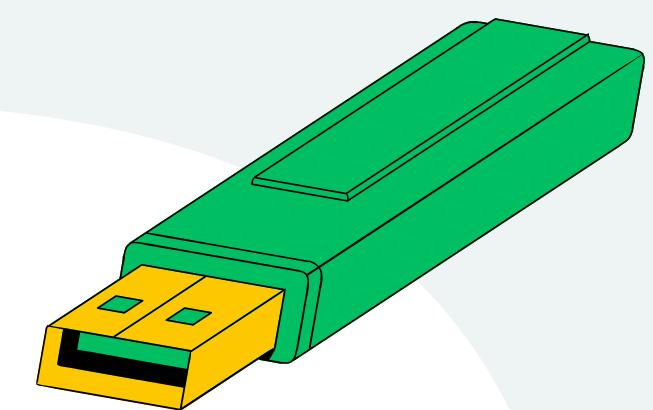
สโรชา เคนเวียง 6720422031



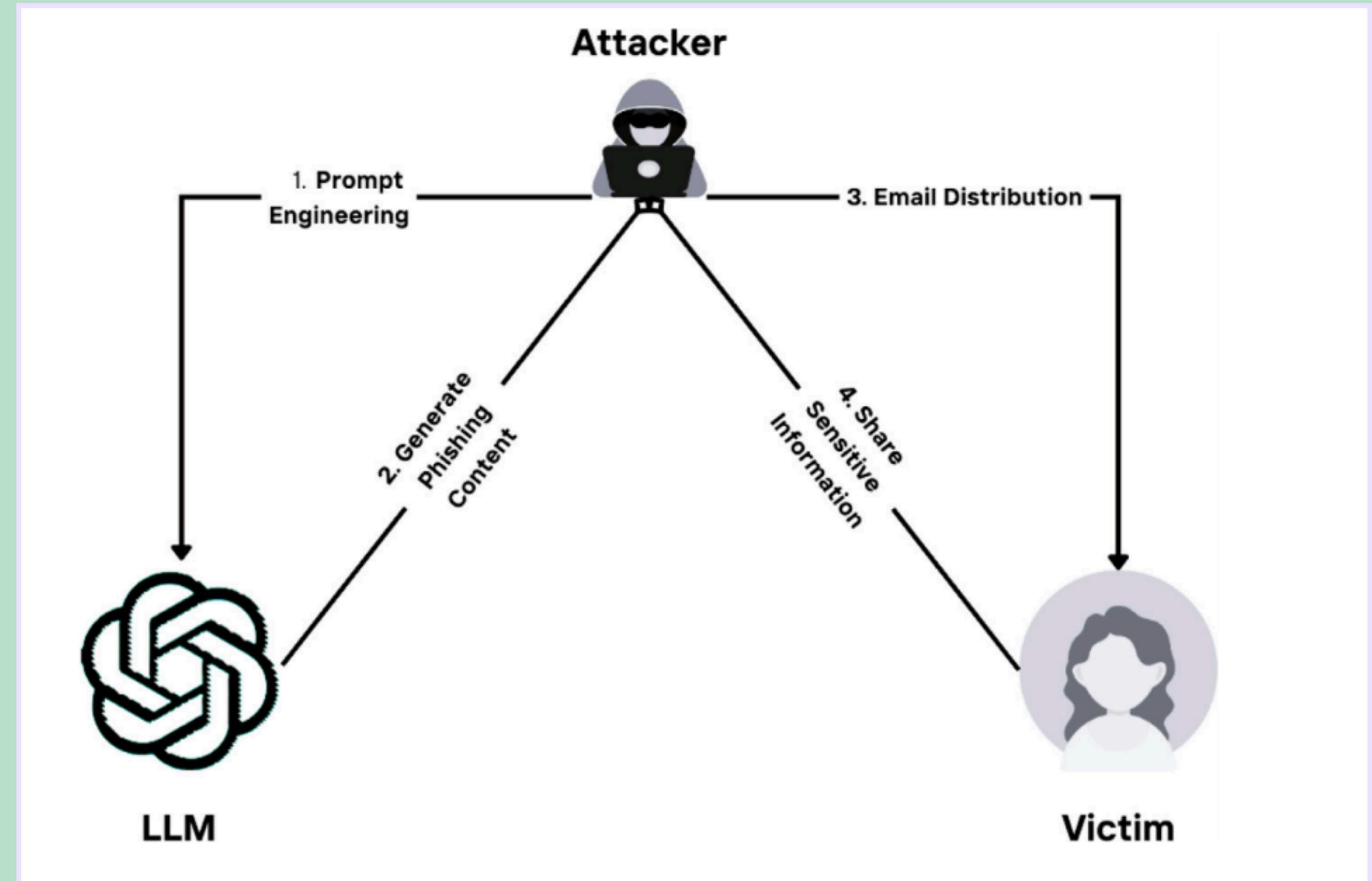
AGENDA



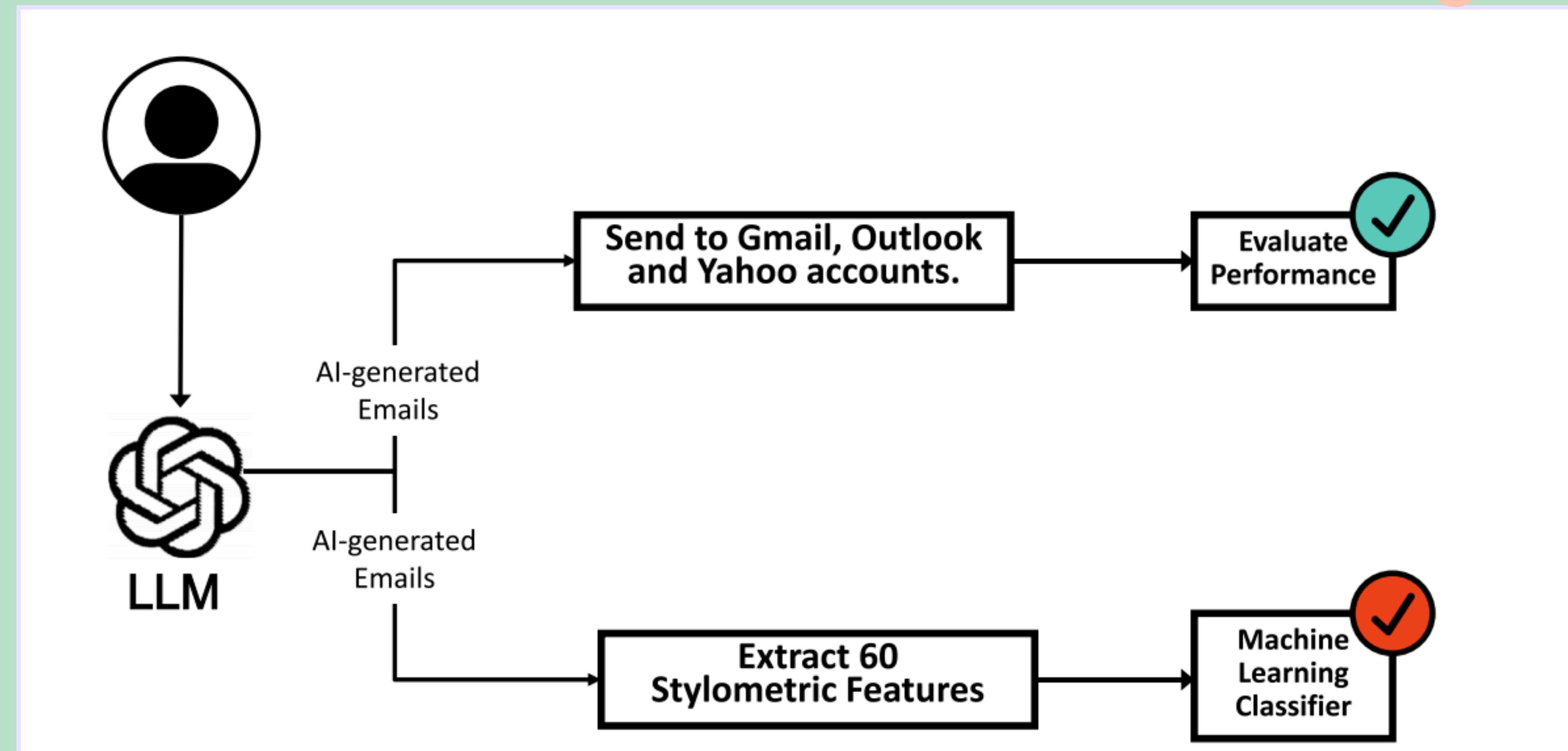
- Introduction and Research Background 01
- Pipeline Machine learning : Reproduce
 - Dataset Processing and Feature Preprocessing 04
 - Modeling & Experiments (Reproduce)
 - Feature Importance (XGBoost) 07
 - Ablation Study 08
 - Performance Comparison 09
- Additional Modeling (New ideas) 10
 - More model and Features Added 11
 - Experimental Results 13
- Conclusion 14



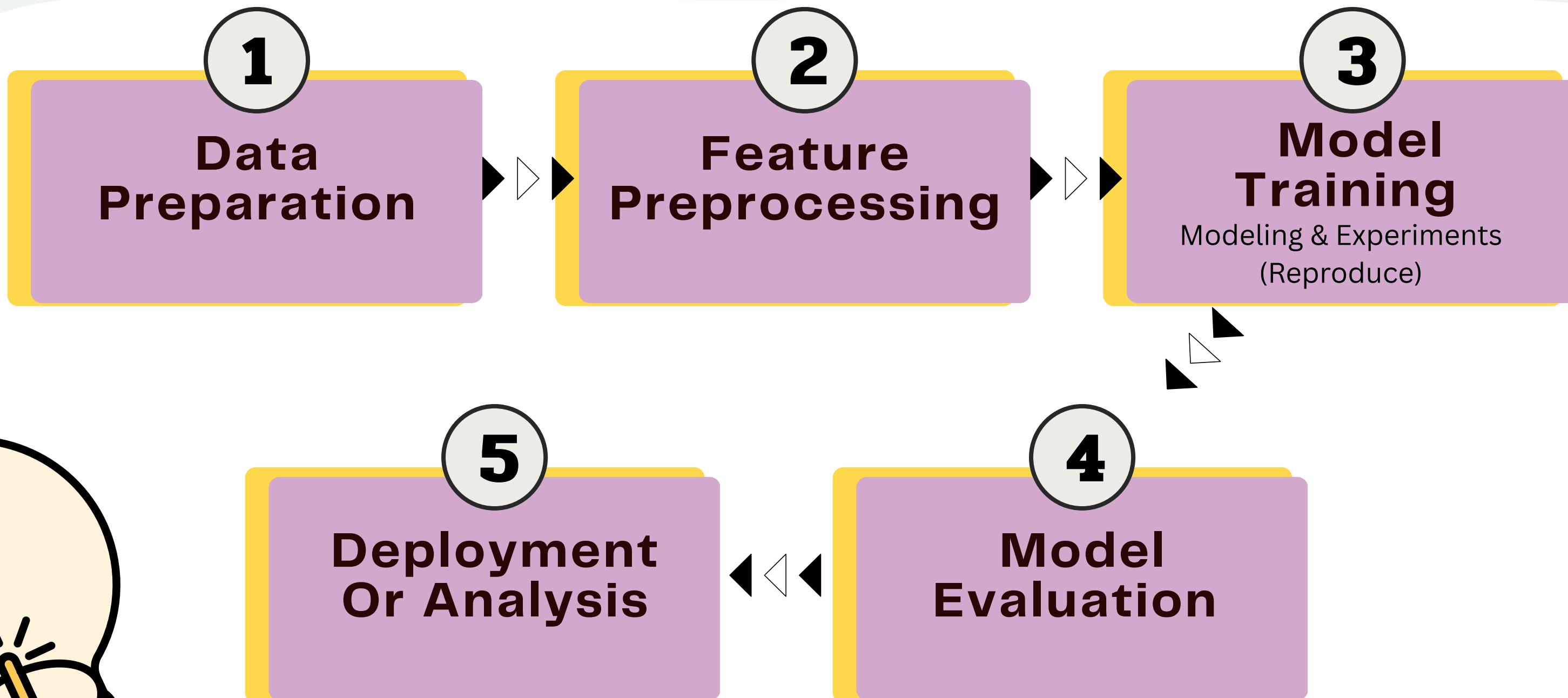
BACKGROUND



OVERVIEW OF THIS STUDY



PIPELINE MACHINE LEARNING



DATA PREPARATION AND PREPROCESSING

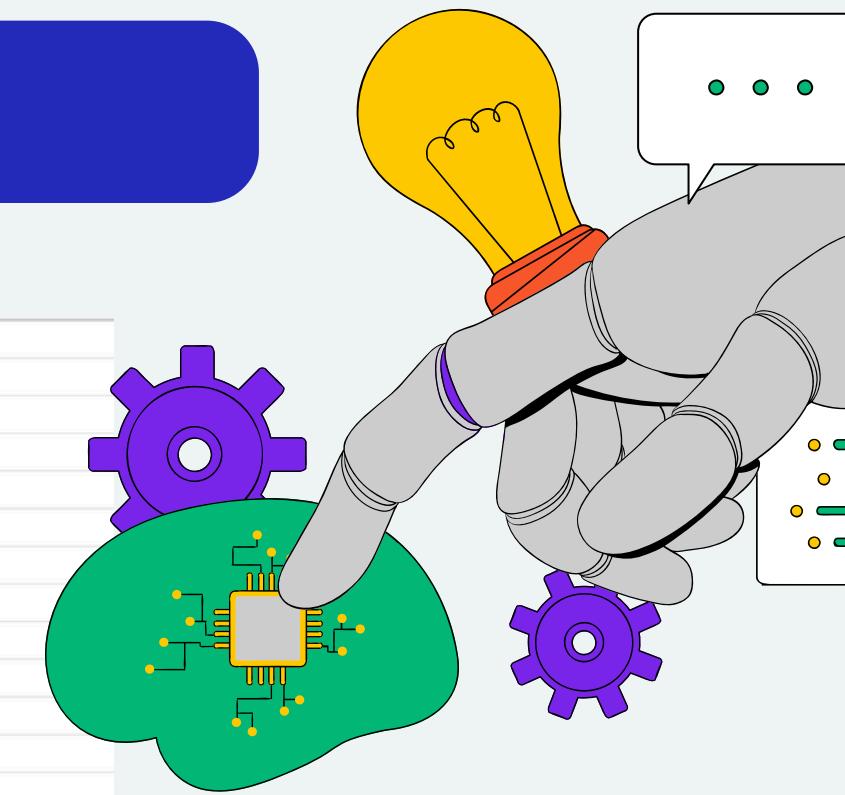
1. Dataset (Train–Test)

Subject	Body	Label
An Exciting Opportunity to Engage in a Per Dear David,	Dear	Legitimate
New Exciting Relationship Opportunity! Fr Dear Julia,	Dear	Phishing
Important Account Notification Å¢â€œ Dear Toni,	Dear	Legitimate
Exclusive Financial Opportunity Awaits - Ir Dear Chelsea,	Dear	Phishing
Introducing Groundbreaking Health Enhanc Dear Christine,	Dear	Legitimate
Urgent Security Alert: Action Required for From: TechZPro Security	Dear	Phishing
Urgent Action Required: Strange Activities Dear Robert,	Dear	Phishing
Security Alert & Account Verification Nee Dear Melanie,	Dear	Phishing
Exciting Opportunity for a Personal Relatic Hello Jonathan,	Hello	Legitimate
Exciting Job Opportunity with Flexible Wor Dear Katherine,	Dear	Legitimate
IMPORTANT: Immediate Response Requir Dear Alfred,	Dear	Phishing
Exciting Opportunity Awaits You - Verify Yc Dear Kimberly,	Dear	Phishing
Exciting New Special Offer Just For You, M Dear Marcus,	Dear	Legitimate
Important Technical Update - Action Requ Dear Samantha,	Dear	Legitimate
Important Notice about Your MySecureBa Dear Amanda,	Dear	Phishing

TRAIN.CSV

Subject	Body	Label
Exclusive Personal Relationship Opp Dear Sean,	Dear	Legitimate
Exceptional Work-From-Home Job Oj Dear Christopher,	Dear	Phishing
Notification of Recent Account Activi Dear Kristen,	Dear	Legitimate
Exciting Work-from-Home Job Opport Sender: HR Dept.	Sender:	Phishing
Exclusive Promotional Offer Just for Y Dear Lindsay,	Dear	Phishing
Exclusive Promotional Offer from "Dy Dear Kyle,	Dear	Phishing
Exciting Job Opportunity Awaits You a Dear Elizabeth,	Dear	Legitimate
Fantastic Work-From-Home Opportu Dear John,	Dear	Phishing
Urgent Update Required for Your Hea From: service@healmngtsolution.us	From:	Phishing
Exciting Job Opportunity and Immedi [Acme Corp Logo]	[Acme Corp Logo]	Phishing
Immediate Job Opportunity at Tech S Dear Mark,	Dear	Phishing
Important Technical Update - Securit Dear Andrew,	Dear	Phishing
Important Account Notification - Plea Dear Rebecca,	Dear	Legitimate
New Opportunity to Connect: Special Dear Mike,	Dear	Legitimate

TEST.CSV



2. รวมคอลัมน์ Subject and Body เข้าด้วยกัน เป็น combined_text

```
# สร้างคอลัมน์ combined_text สำหรับการ extract features
if has_subject and has_body:
    train_df['combined_text'] = train_df['Subject'].astype(str) + " " + train_df['Body'].astype(str)
    test_df['combined_text'] = test_df['Subject'].astype(str) + " " + test_df['Body'].astype(str)
    print("✓ สร้าง combined_text จาก Subject และ Body")
else:
    text_column = train_df.columns[0]
    print(f"⚠️ ไม่คอลัมน์ '{text_column}' เป็นข้อความ")
    train_df['combined_text'] = train_df[text_column].astype(str)
    test_df['combined_text'] = test_df[text_column].astype(str)
```

4. Feature Normalization (StandardScaler)

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

3. Missing Value เติมค่าที่หายไปด้วยค่าเฉลี่ย

```
# เติมค่าที่หายไปด้วยค่าเฉลี่ย
X_train = X_train.fillna(X_train.mean())
X_test = X_test.fillna(X_test.mean())
```

5. การแปลง Label ของข้อมูลเป็นตัวเลข

```
# Encode target labels
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)

class_names = label_encoder.classes_
print(f"คลาส: {class_names}")
```

MODELING & EXPERIMENTS (REPRODUCE)

4 Models Used

1. Logistic Regression
2. Support Vector Machine (SVM)
3. Random Forest
4. XGBoost

Feature Importance +
Ablation Study

Evaluation Metrics

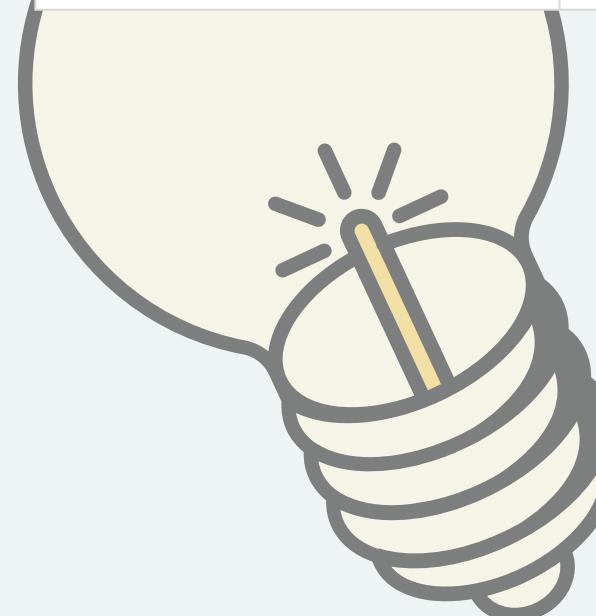
- Accuracy
- Precision
- Recall
- F1-Score
- AUC



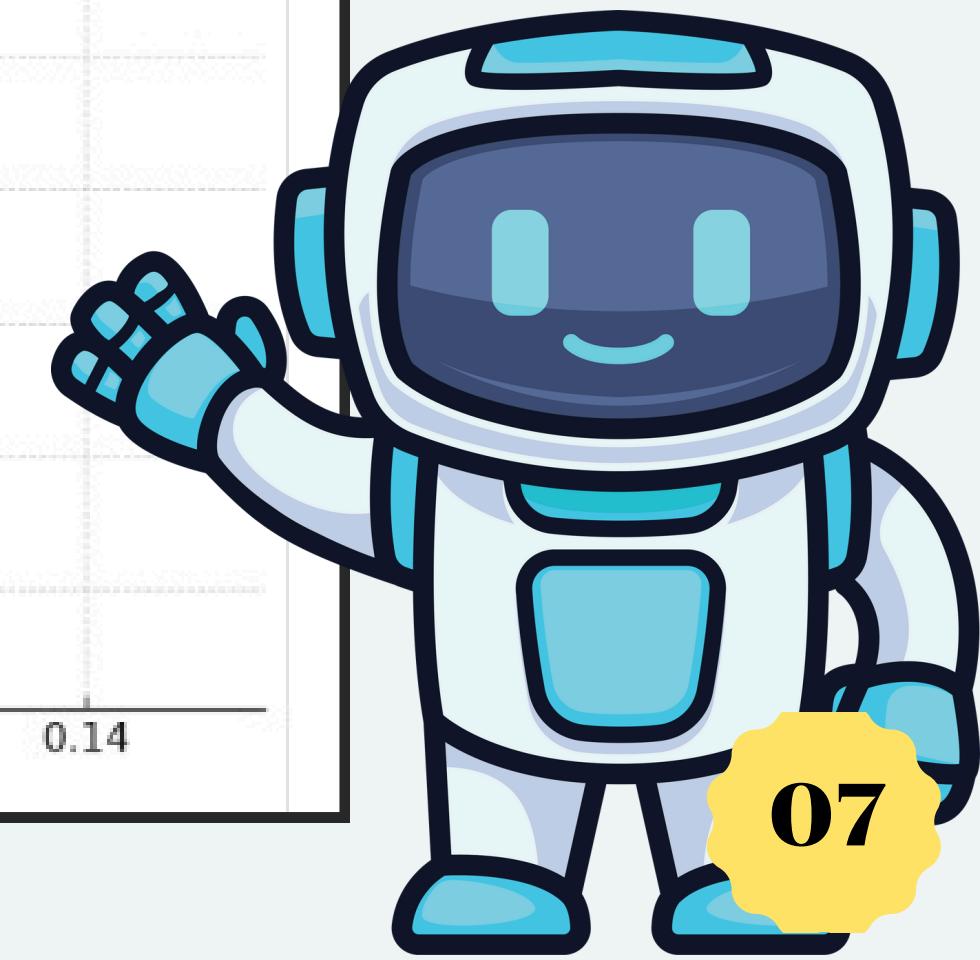
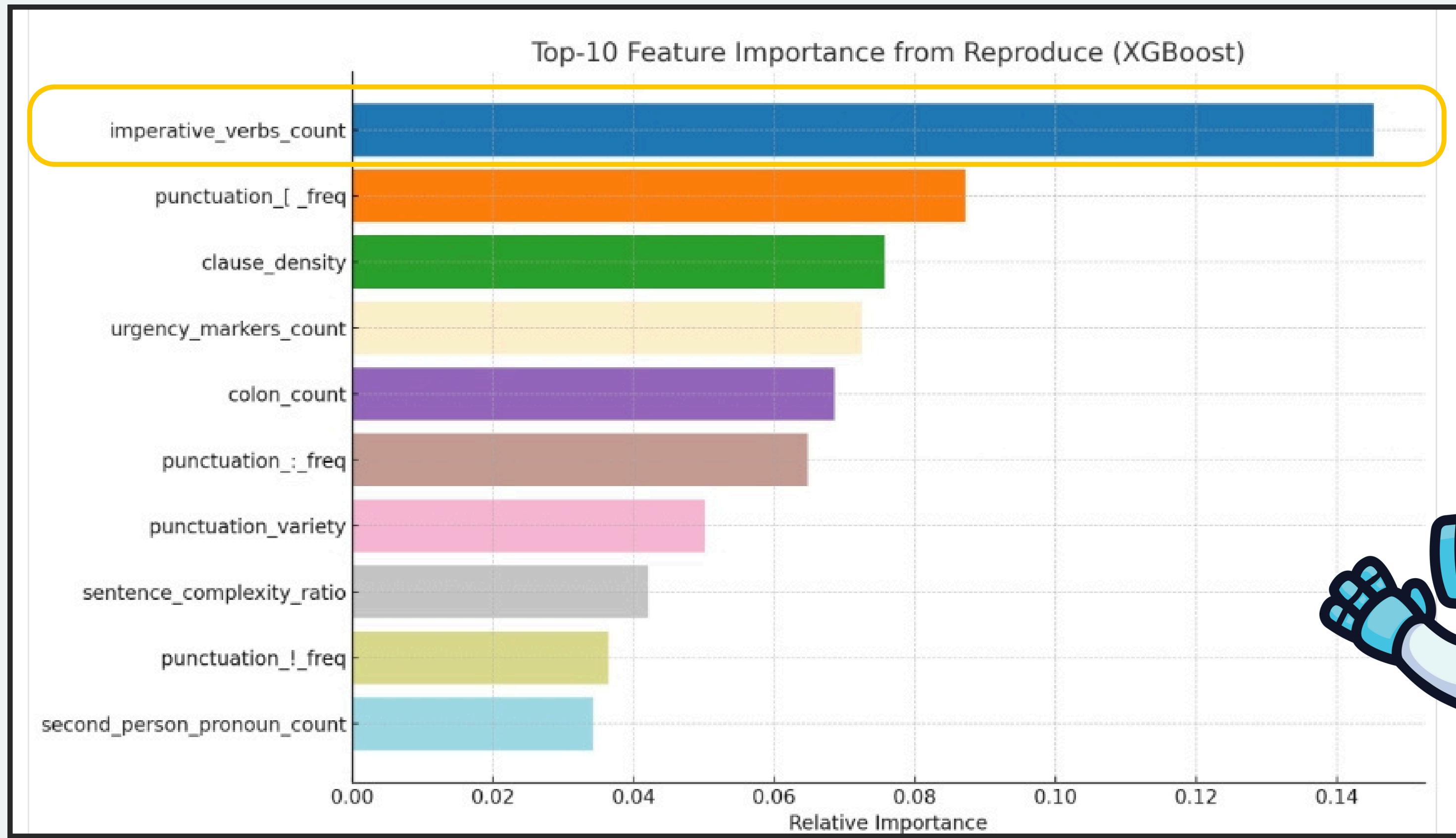
MODELING & EXPERIMENTS (REPRODUCE) - RESULTS

Experimental Results : All of Model

Algorithm	Accuracy			Precision			Recall			F1-Score			AUC-Score		
	Modeling	Our	Paper	Diff	Our	Paper	Diff	Our	Paper	Diff	Our	Paper	Diff	Our	Paper
XGBoost	0.9231	0.96	-0.04	0.9333	0.96	-0.03	0.9231	0.96	-0.04	0.9226	0.96	-0.04	0.9704	0.99	-0.02
Logistic Regression	0.6923	0.92	-0.23	0.7124	0.93	-0.22	0.6923	0.92	-0.23	0.6848	0.92	-0.24	0.8875	0.98	-0.09
Random Forest	0.8846	0.92	-0.04	0.8869	0.93	-0.04	0.8846	0.92	-0.04	0.8844	0.92	-0.04	0.9704	0.99	-0.02
SVM	0.8077	0.85	-0.04	0.825	0.88	-0.06	0.8077	0.85	-0.04	0.8051	0.84	-0.03	0.9586	0.96	0



FEATURE IMPORTANCE (XGBOOST) REPRODUCE



ABLATION STUDY

ตรวจสอบฟีเจอร์กลุ่มใด “จำเป็นจริง ๆ” ต่อความแม่นยำของโมเดล

Ablation Study Results:

Evaluation Metrics	Accuracy	Precision	Recall	F1-Score	Number Features
All Features	0.9231	0.9333	0.9231	0.9226	61
Without Top 10	0.8077	0.825	0.8077	0.8051	51
Without Top 20	0.9231	0.9333	0.9231	0.9226	41
Without Top 30	0.8462	0.8545	0.8462	0.8452	31

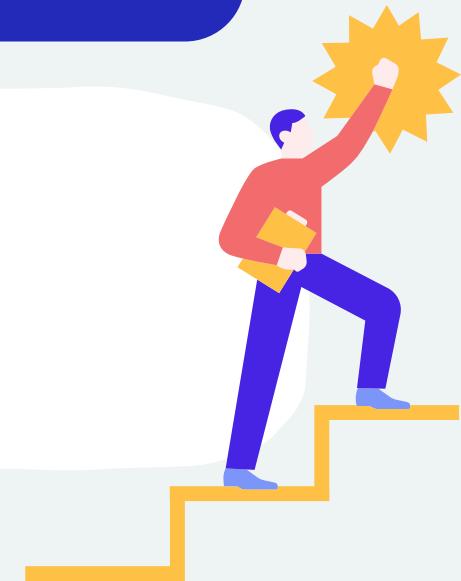
Top 10 คือ ฟีเจอร์กลุ่มนี้คือ กลุ่มหลัก ที่ช่วยแยก phishing vs legitimate

Top 30 features → Accuracyลดลงอีกครึ่ง ฟีเจอร์อันดับ 21–30 คือกลุ่ม “supportive signal”
ไม่ใช่ฟีเจอร์หลัก แต่ช่วยเสริมการแยกประเภท

ADDITIONAL MODELING (NEW IDEAS)

Assumption :

1. การเพิ่มฟีเจอร์ช่วยเพิ่มความแม่นยำในการคำนวณ
2. นอกจาก 4 โมเดลที่กล่าวถึงแล้ว ยังมีโมเดลใดบ้างที่สามารถให้ความแม่นยำในการคำนวณที่สูงกว่า?



7 Models Used

1. L1 Regularization (SGD)
2. L2 Regularization (SGD)
3. Elastic Net (SGD)
4. K-Nearest Neighbors
5. Decision Tree
6. Neural Network
7. Gradient Boosting

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- AUC

+Additional Feature

ADDITIONAL MODELING (NEW IDEAS) - MORE MODEL

01

L1 Regularization (Lasso)

- บังคับให้ค่าบิวตี้系数บางตัวเป็น 0
- เลือก feature selection อัตโนมัติ
- เหมาะกับข้อมูลที่มีฟีเจอร์จำนวนมาก
- โมเดลเรียบง่ายและต่ความได้ดี
- อาจเกิด underfitting และไวต่อ noise

02

L2 Regularization (Ridge)

- ไม่เซ็ตบิวตี้系数เป็น 0 แต่ลดขนาดบิวตี้系数ทั้งหมดอย่างสม่ำเสมอ
- ช่วยเพิ่มเสถียรภาพของโมเดล และลด overfitting เสถียรและก้าวไปได้ดี
- ไม่สามารถตัดฟีเจอร์ที่ไม่สำคัญได้

03

Elastic Net (SGD)

- ผสมระหว่าง L1 และ L2 เพื่อให้ได้ทั้งผลของการ feature selection และความเสถียรของบิวตี้系数
- เหมาะเมื่อฟีเจอร์มีความสัมพันธ์กัน
- สมดุลระหว่าง L1 และ L2
- ต้องปรับ hyperparameter มากกว่าแบบเดียว

04

K-Nearest Neighbors(KNN)

- โดยคำนึงจากตำแหน่งของจุดใหม่ แล้วดู k ตัวใกล้ที่สุด ของ training data
- เข้าใจง่าย, สามารถเรียนรู้ความสัมพันธ์แบบไม่เชิงเส้น
- คำนึงซ้ำ, มีปัญหาเมื่อจำนวนฟีเจอร์สูง

05

Decision Tree

- จำแนกข้อมูลด้วยการสร้างกฎแบบลำดับขั้นที่เลือกตัวแปรแบ่งกลุ่มเพื่อลดความไม่บริสุทธิ์ของข้อมูล
- ต่ความง่ายและรองรับความสัมพันธ์ไม่เชิงเส้น
- เกิด Overfit หากไม่ควบคุมความลึก

06

Neural Network

- โมเดลแบบหลายชั้นที่เรียนรู้ความสัมพันธ์ไม่เชิงเส้นผ่านการคำนวณของ neuron หลายชั้น
- เรียนรู้ pattern ซับซ้อนได้ดี
- ใช้เวลาเทรนมากและต่ความยาก

07

Gradient Boosting

- วิธีรวมโมเดลอ่อนหลายตัวแบบลำดับ โดยให้แต่ละตัวไม้อธิบายส่วนที่โมเดลก่อนหน้าคำนายนิด
- ประสิทธิภาพสูงและลด bias ได้ดี
- ต้องปรับแต่งพารามิเตอร์และอาจใช้เวลาฝึกนาน

10

ADDITIONAL MODELING (NEW IDEAS)



01

Emotional Tone Score - คะแนน
โทนอารมณ์

positive_emotion_words

Example : 'great', 'good', 'excellent',
'amazing', 'wonderful', 'happy', 'nice'

negative_emotion_words

Example : 'bad', 'terrible', 'awful',
'horrible', 'sad', 'angry', 'frustrated'

02

Formality Index
- ดัชนีความเป็นทางการ

Example : 'sincerely', 'regards',
'respectfully', 'cordially', 'yours'

03

Legal Terminology Density
- ความหนาแน่นคำทางกฎหมาย

Example : 'terms', 'conditions', 'agreement',
'policy', 'compliance', 'regulation'

04

Sentence Structure Complexity - ความซับซ้อน
โครงสร้างประโยค

คำนวณความแปรปรวนของความยาว
ประโยค

05

Question Mark Density -
ความหนาแน่นเครื่องหมายคำถาม

การใช้เครื่องหมายคำถามที่มากเกินไป ซึ่งอาจเป็น
ลักษณะการข้อจุ่ง

Additional Modeling (New ideas)

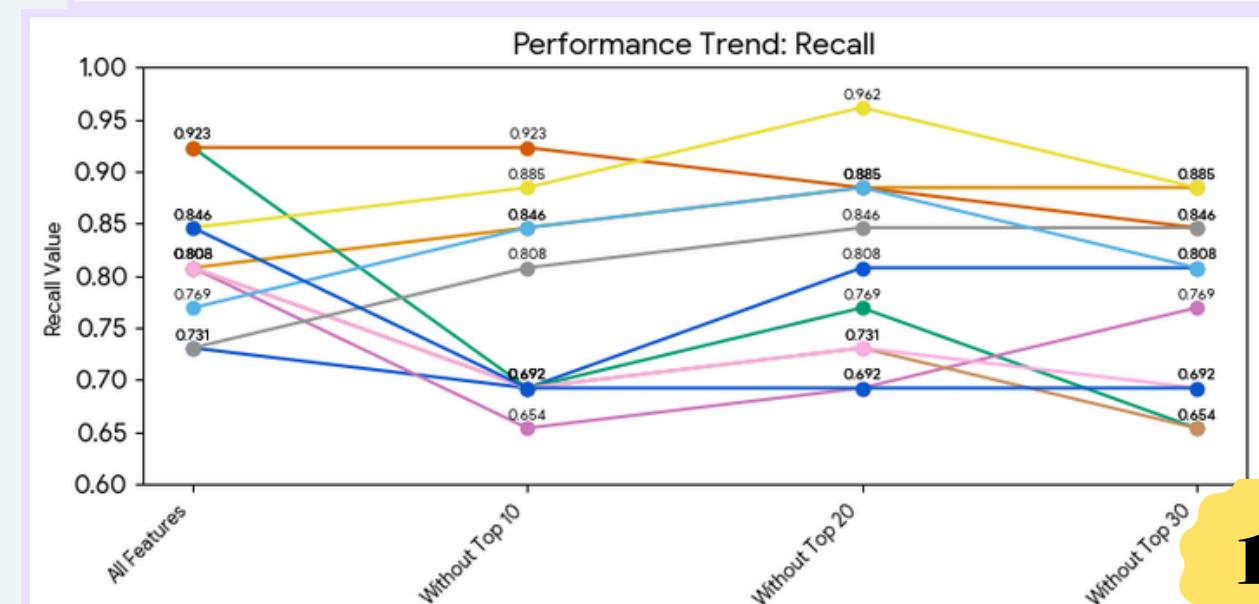
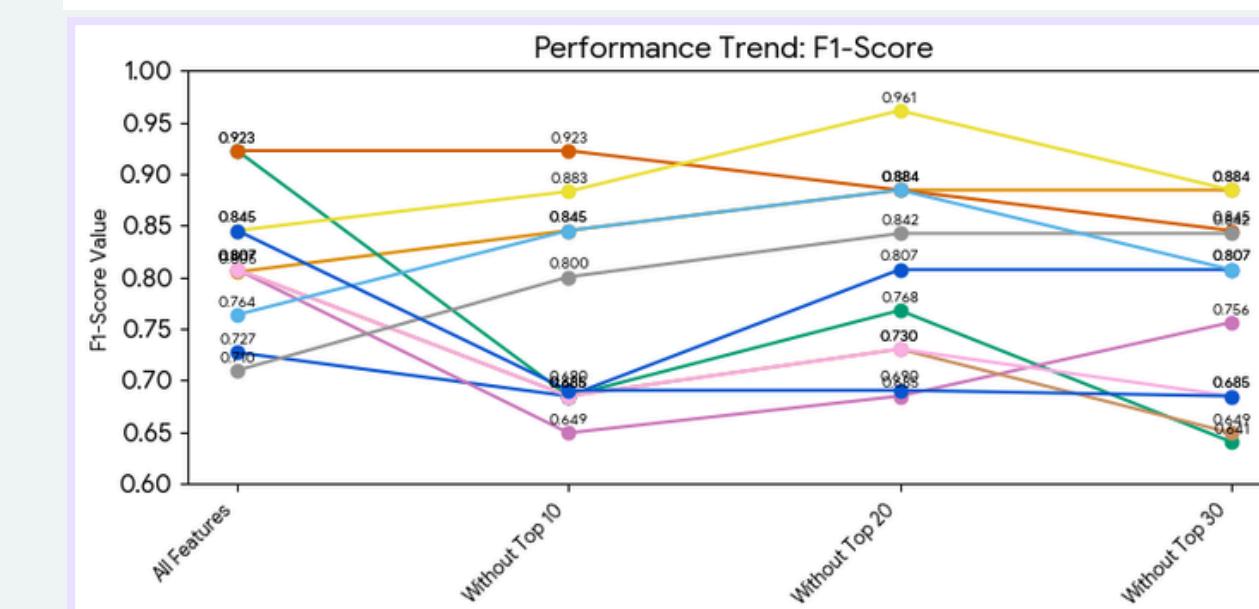
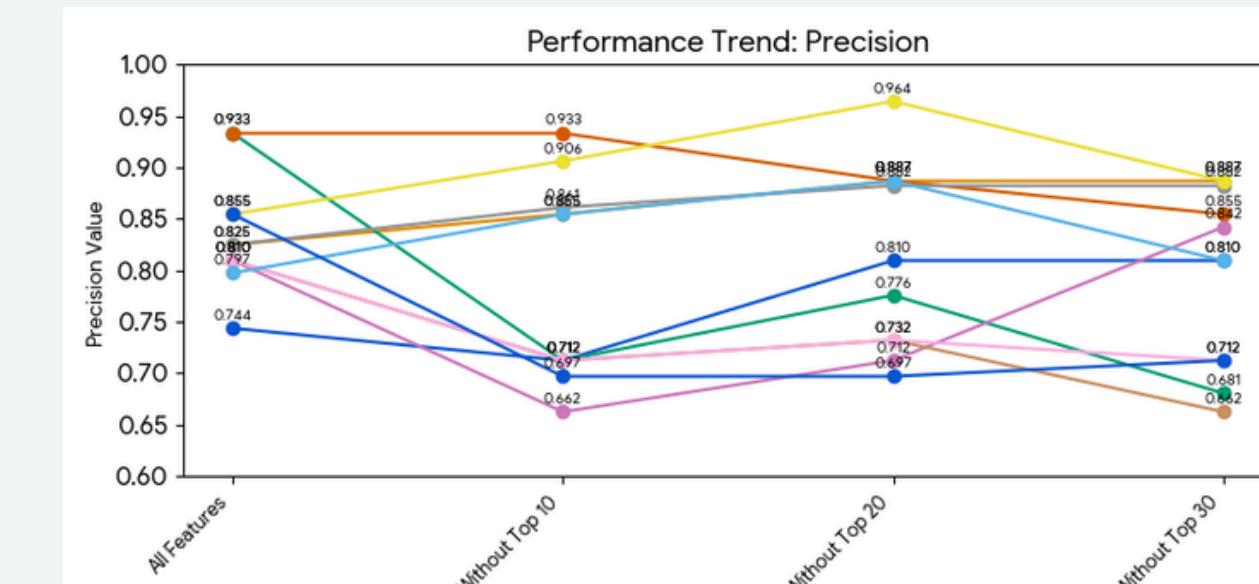
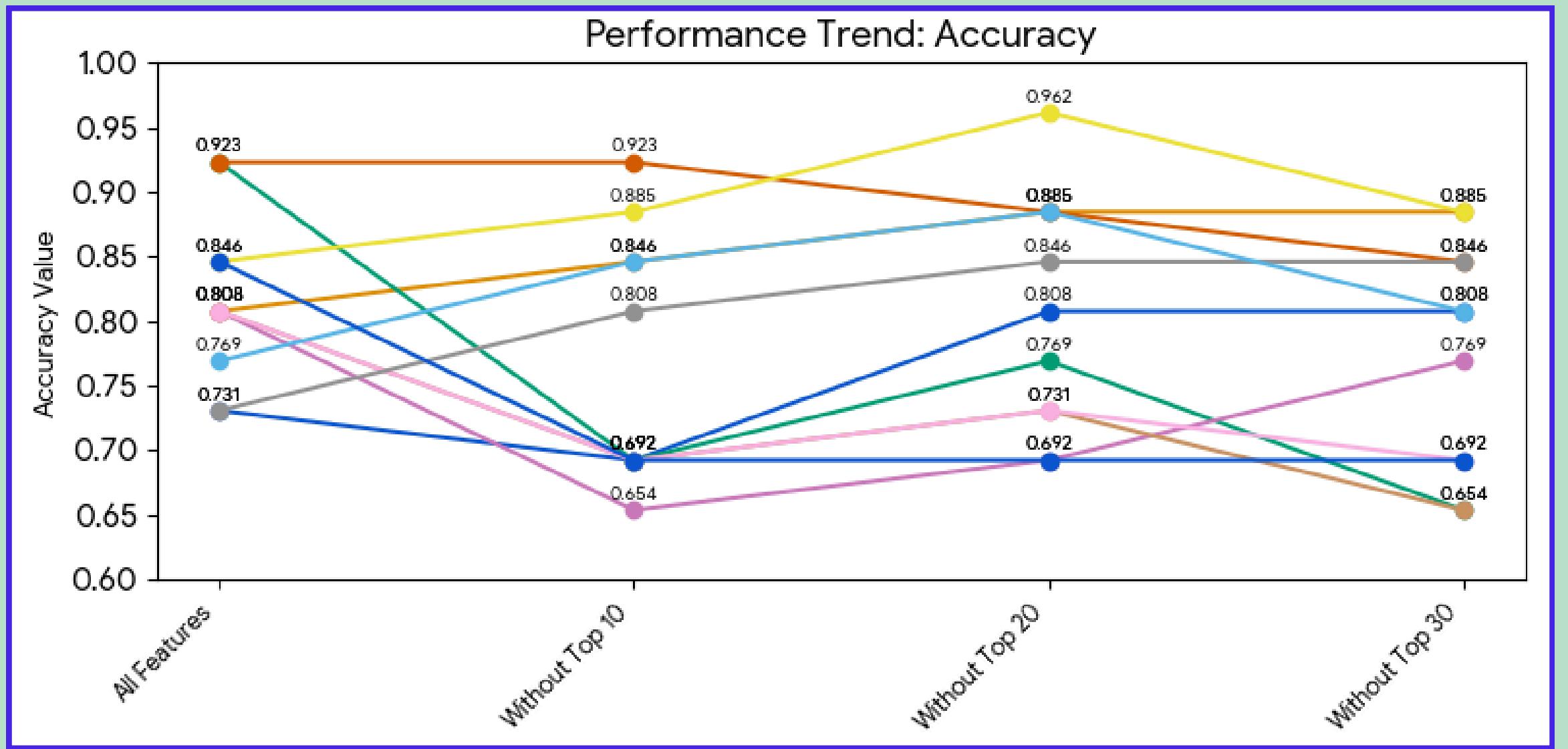
Experimental Results : All of

Modeling	Accuracy	Precision	Recall	F1-Score	AUC-Score
Random Forest	0.9231	0.9333	0.9231	0.9226	0.9882
XGBoost	0.9231	0.9333	0.9231	0.9226	0.9704
Decision Tree	0.8462	0.8545	0.8462	0.8452	0.8462
Gradient Boosting	0.8462	0.8545	0.8462	0.8452	0.9586
SVM	0.8077	0.825	0.8077	0.8051	0.9586
Elastic Net (SGD)	0.8077	0.8095	0.8077	0.8074	0.8935
L1 Regularization (SGD)	0.8077	0.8095	0.8077	0.8074	0.9053
L2 Regularization (SGD)	0.8077	0.8095	0.8077	0.8074	0.9112
Neural Network	0.7692	0.7974	0.7692	0.7636	0.929
Logistic Regression	0.7308	0.7437	0.7308	0.7271	0.8994
K-Nearest Neighbors (KNN)	0.7308	0.825	0.7308	0.7271	0.8787

ADDITIONAL MODELING (NEW IDEAS) -ABLATION STUDY

Note

Model	XGBoost
Logistic Regression	●
SVM	●
Random Forest	●
Elastic Net (SGD)	●
Gradient Boosting	●



- **XGBoost** เป็นโมเดลที่ค่าความแม่นยำไม่ได้เปลี่ยนแปลงมากเมื่อมีการลดฟีเจอร์ **แต่ Random Forest** ค่า Accuracy ลดลงอย่างมาก จาก 92% → 69% การลด feature ลงมีผลอย่างมาก
- ส่วนโมเดลอื่น เช่น Logistic Regression, Neural Network, KNN มีประสิทธิภาพดีลงเมื่อลดจำนวน Feature ลง เพราะ อาศัยจำนวนฟีเจอร์และความสมบูรณ์ของข้อมูลสูง

Summary

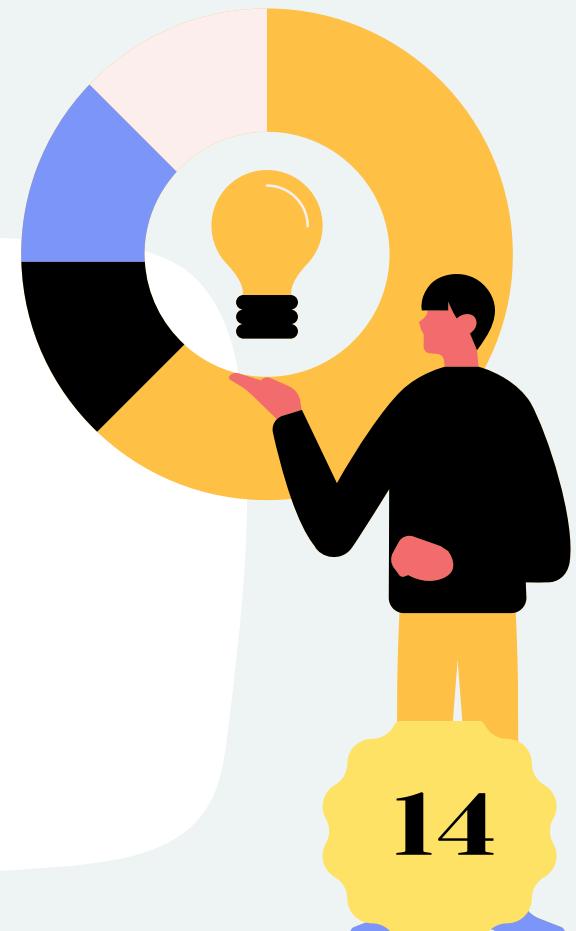
Reproduce

จาก Paper สรุปว่า XGBoost เป็น Model ที่ได้ค่า Accuracy เท่ากับ 96% แม่นยำที่สุด และจากการ Reproduce เราได้ค่า Accuracy เท่ากับ 92.3% เราสมมุติฐานจากการ Reproduce จากค่าที่แตกต่างกันว่า

- ขั้นตอนการทดลอง Preprocess ที่แตกต่างกัน
- ขั้นตอนการเขียน Coding แตกต่างกัน เช่น การ Tuning Hyperparameter ของแต่ละ Model ไม่เท่ากัน

Additional Model

- ผลจากการเพิ่ม Sylometric Feature 5 ตัว และเปรียบเทียบ Accuracy ของแต่ละ Model พบว่าแต่ละ Model มีค่า Accuracy อยู่ในช่วง 73%-92.3% โดยที่ค่า Random Forest และ XG Boost ได้ค่า Accuracy , Precision, Recall, F1-Score ที่เท่ากัน แต่แตกต่างกันที่ค่า AUC ซึ่ง Random Forest ได้มากกว่าที่ 98.8%



THANK YOU

