

Artificial Intelligence for Cybersecurity

Week 5: AI Deployment in Cybersecurity

James Hu

Ph.D. Candidate

Acknowledgements: Dr. Hsinchun Chen



Online M.S. in
AI for Business

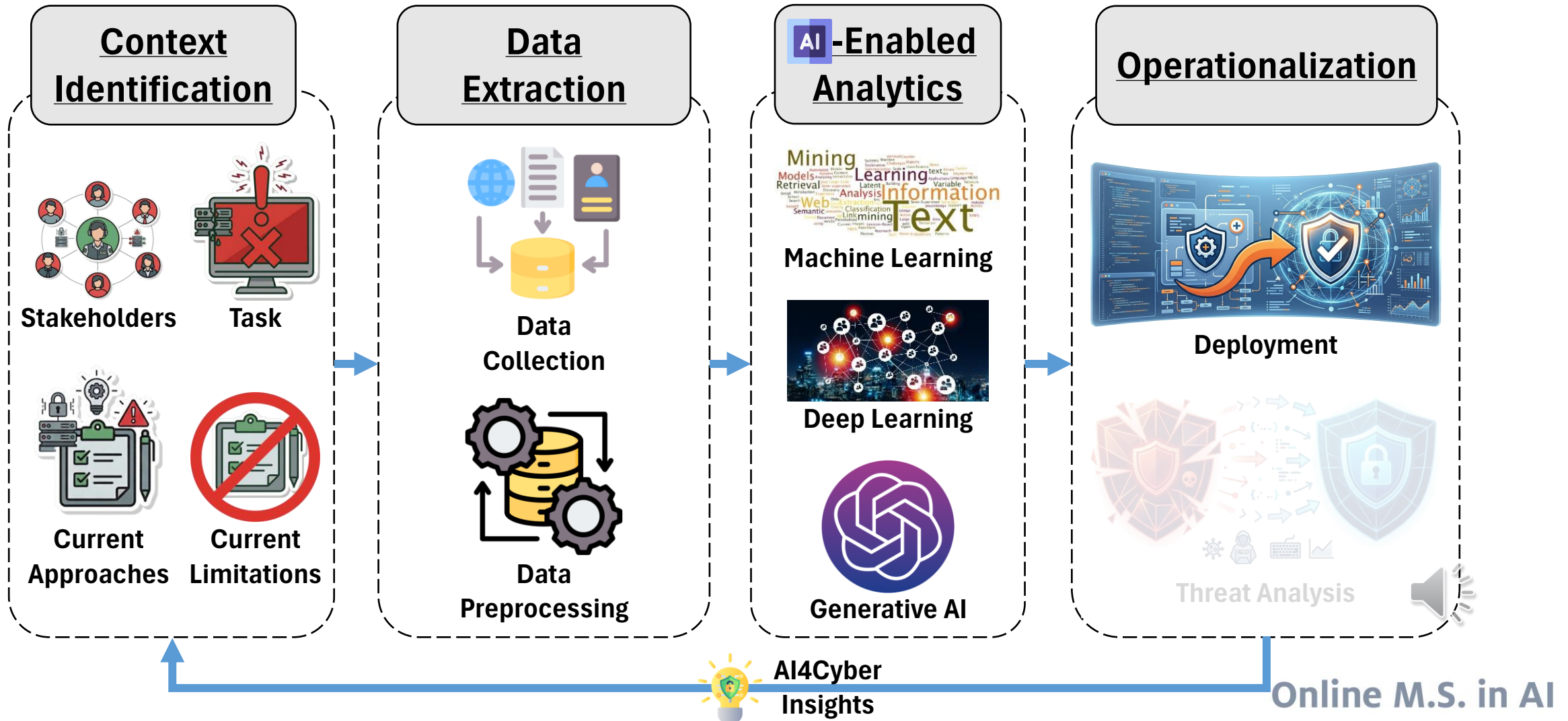


Today's Learning Objectives

- ▶ Understand why holdout datasets and case studies are important for demonstrating an AI-enabled analytic's deployment performance.
- ▶ Identify the importance of the context of “Phishing” and how this can be applied to other cybersecurity areas of your interests.
- ▶ Explain the differences between detecting phishing emails vs detecting vishing attacks and how these differences motivate the choice in AI-enabled analytics



From Week 1: AI4Cyber Lifecycle



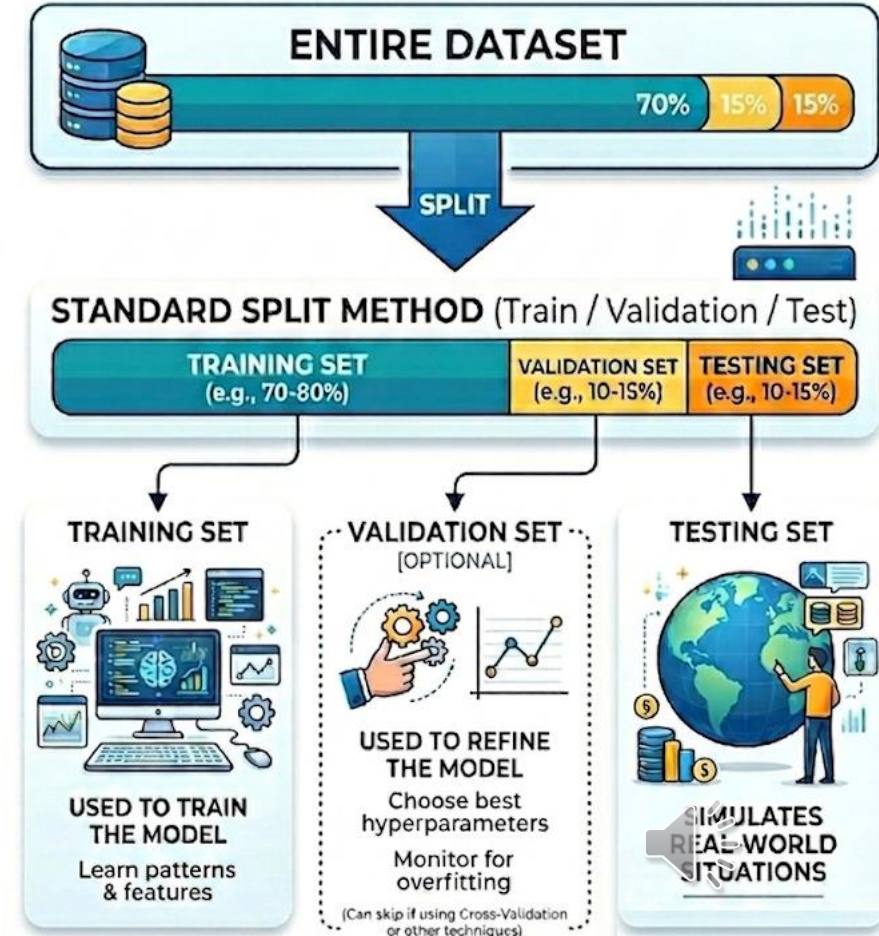
Operationalization: Deployment

- ▶ The final step in the AI4Cyber lifecycle is operationalization, where different projects either focus on deployment or threat analysis.
 - ▶ This week, we will focus on deployment studies, using phishing as an example use-case.
 - ▶ Weeks 6 and 7 will be dedicated to threat analysis, which requires a specialized suite of methods.
- ▶ Deployment refers to the implementation and evaluation of AI-enabled analytics outside of their development environment.
- ▶ Two common strategies for deployment:
 - ▶ Holdout datasets
 - ▶ Case studies



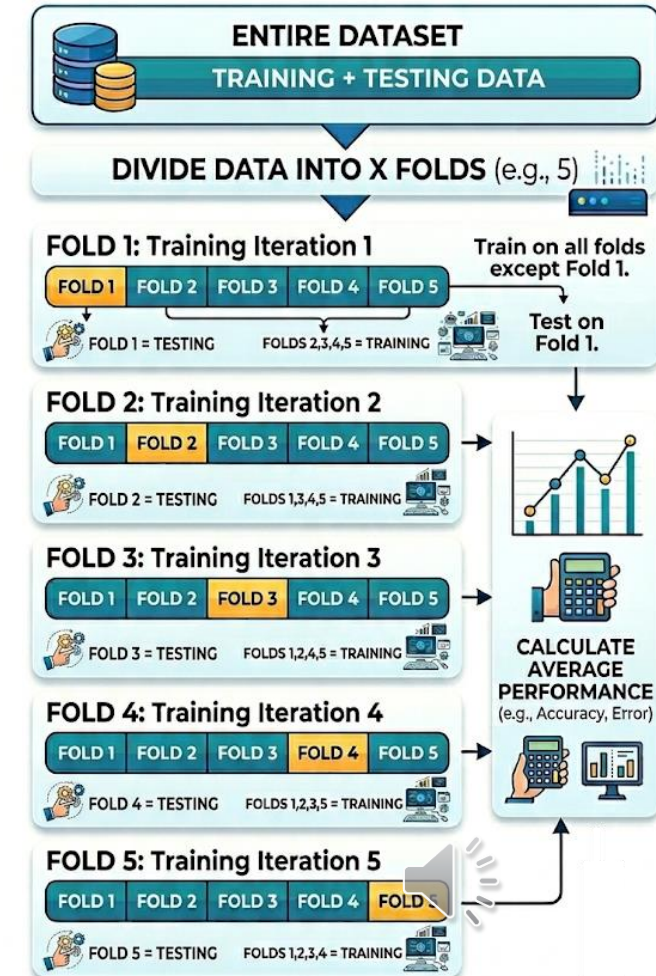
Deployment Strategy: Holdout Datasets

- ▶ A common deployment strategy is using holdout datasets, or the splitting of a dataset into distinct subsets:
 - ▶ **Training Set:** Datapoints that the AI-enabled analytic input during its ML training loop.
 - ▶ **Validation Set:** Datapoints not inputted to the model during training that the authors use to adjust hyperparameters
 - ▶ E.g., Learning rate, model architecture, etc.
 - ▶ **Test Set:** Datapoints completely hidden from AI-enabled analytic during its development process to simulate real-world situations.



Deployment Strategy: Holdout Datasets

- ▶ Splitting small datasets may introduce bias in your results.
 - ▶ **Consider:** Assigning all the “easy questions” to the testing set would artificially increase a model’s deployment “performance.”
- ▶ *X*-Fold Cross Validation removes bias via five steps:
 1. Divide dataset into *X* subsets (folds)
 2. Assign one fold to the testing set. The rest are assigned to the training set.
 3. Train and evaluate using the training and testing set.
 4. Repeat Steps 2 and 3 until all folds were assigned to the testing set once.
 5. Average all metrics across each fold.



Deployment Strategy: Case Studies

- ▶ Another common deployment strategy is case studies, or the construction of separate environments to evaluate developed AI-enabled analytics in.
 - ▶ Case study environments should closely represent how the AI-enabled analytic is expected to be used by the key stakeholders.
- ▶ Case studies vary wildly between each AI4Cyber project, but typically involve:
 - ▶ **Situated Implementations (Low Risk):** AI-enabled analytic is evaluated in a simulation that closely emulates production environments.
 - ▶ Often the preferred case study type for academic and immature industry projects.
 - ▶ **Real-World Implementations (High Risk):** AI-enabled analytic is directly implemented in a production environment with actual customers and end-users.
 - ▶ Only reserved for mature industry projects but yields real usage analytics through pilot studies.
 - ▶ Real-world implementation case studies can often be standalone AI4Cyber projects.





Artificial Intelligence for Cybersecurity

Week 5: Operationalized Deployment

James Hu

Ph.D. Candidate

Acknowledgements: Dr. Hsinchun Chen



Online M.S. in
AI for Business



Background: What is Phishing?

- ▶ Suppose you are an employee at a government agency. You have access to sensitive information, and your machine is connected to the wider agency network.
- ▶ An adversary would like to access your sensitive information or the agency network. How would they do so?
 - ▶ Common approach, exploit the weakest link in the system (you, the human) and induce human mistakes.
- ▶ Phishing is a form of social engineering where adversaries deceive humans into divulging private information or access.
 - ▶ Today, phishing is typically done via a digital medium.

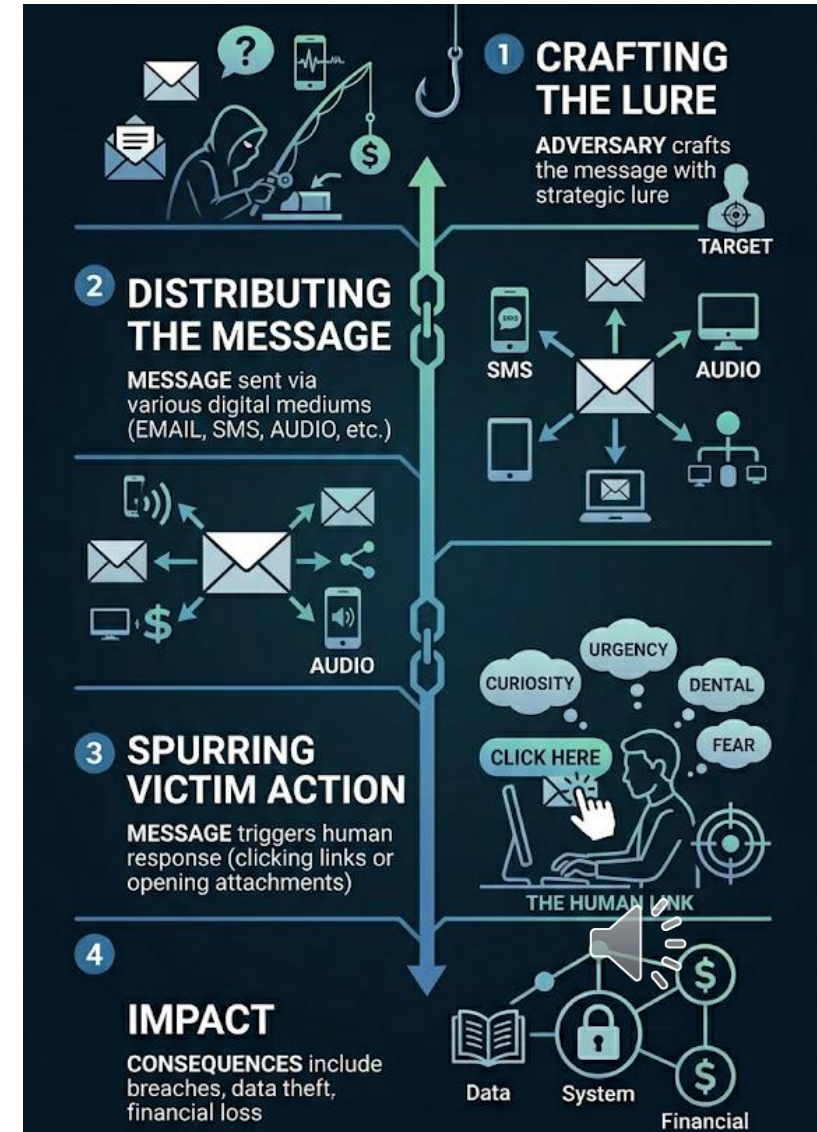


Background: What is Phishing?

- ▶ A modern phishing attack typically operates in four general steps:

1. The adversary crafts a message containing a "lure" to spur the victim into action.
2. The message is sent via various digital mediums.
 - ▶ Includes email (most common), SMS, and audio calls.
3. The message spurs the victim into action, which can include the victim:
 - ▶ Clicking a malicious link, which directs the victim to a phishing website to steal their credentials.
 - ▶ Opening a malicious attachment, which installs malware or ransomware directly onto the victim's device.
4. The adversary uses the stolen data for identity theft, financial fraud, or the access as a foothold for a larger data breach.

Cybersecurity analysts aim to disrupt phishing attacks at or before this step



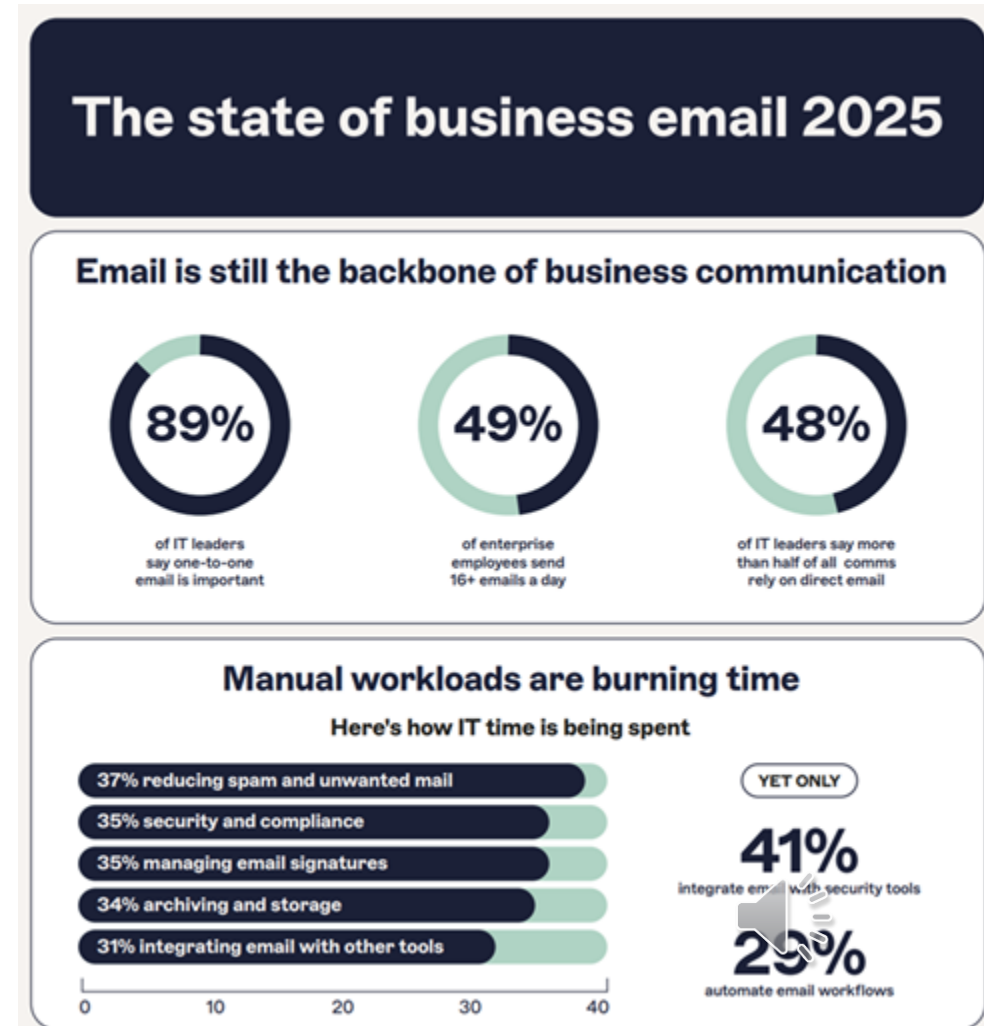
Background: Why care about Phishing?

- ▶ In 2024, the FBI recorded over 57K successful phishing attacks.
 - ▶ 22.5% of all successful cybercrime reports.
 - ▶ 157 **successful** phishing attacks **per day**.
- ▶ Even one successful attack can lead to severe consequences.
 - ▶ Phishing was reported as the top fraud category in 2023 with a reported **global loss of \$2.7 billion**.
- ▶ Volume and severe consequences of phishing attacks motivate an automated approach for phishing detection.
 - ▶ The AI4Cyber lifecycle can guide the development of AI-enabled analytics in addressing phishing attacks.



Email Phishing: Context Identification

- ▶ Email remains the dominant communication medium for firms.
 - ▶ As such, phishing emails is a popular medium for phishing attacks.
- ▶ Firms are dedicating significant manual efforts towards detecting and preventing phishing emails.
 - ▶ Often the work of Email Security Administrators.



Online M.S. in AI



Email Phishing: Context Identification

- ▶ Firms also implement employee training in a defense-in-depth security architecture against phishing emails.
 - ▶ **Defense-in-Depth:** Layered, holistic cybersecurity strategy with multiple redundant security implementations.
- ▶ To robustify the victims (employees) against phishing attacks, firms train their employees to recognize common phishing characteristics.
 - ▶ Unfortunately, research has repeatedly demonstrated that employee training + human detectors still fail to detect well-crafted phishing emails.



Email Phishing: Context Identification



Stakeholder

- **Primary User:** Email Security Administrators.
- **Critical Asset:** Corporate email accounts and sensitive employee information.



Task: Phishing Detection

- **Objective:** Phishing emails are blocked. Suspicious emails properly labeled to alert employees.
- **Current Data:** Prior normal/phishing emails and their respective features.



Current Approaches

- **Core Question:** Should this email be delivered to the inbox, moved to junk, or stripped of its links?
- **Current Method:** Manual analysis of emails by IT team. Employee training.



Current Limitations

- **Data-Wise:** Textual data contains dependencies that are difficult to model
- **Method-Wise:** Manual effort is not scalable. Employees are inconsistent with their evaluations.



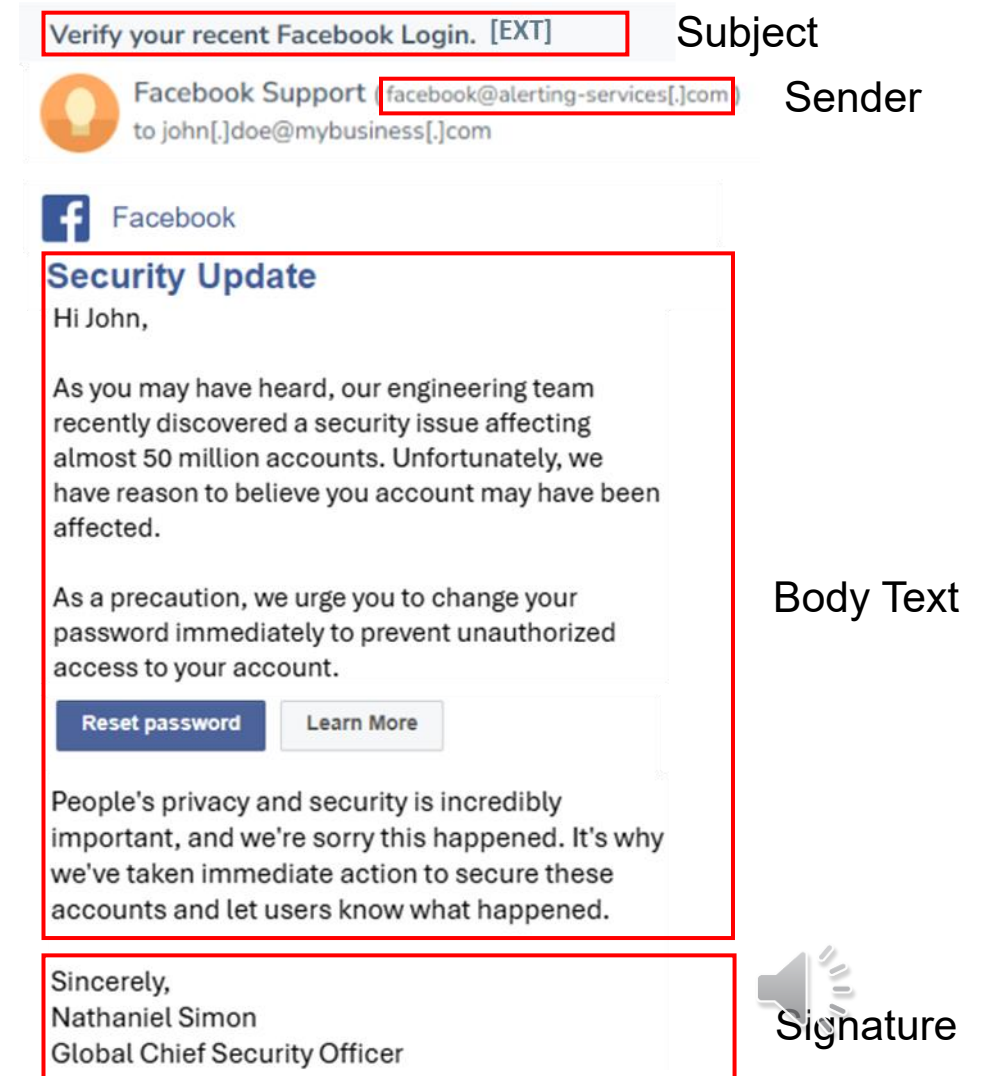
■ Email Phishing: Data Extraction – Collection

- ▶ Firms often keep a detailed internal database of all emails sent and received.
- ▶ Common sources:
 - ▶ Closed-Source
 - ▶ Internal Company Records
 - ▶ Open-Source
 - ▶ Repositories (GitHub, PhishTank, Abuse.ch)
 - ▶ Standard Public Datasets (EnronEmails, Nazario Phishing Corpus)
- ▶ Collection methods: Internal SQL queries, API endpoints, Web crawlers



Email Phishing: Data Extraction – Preprocessing

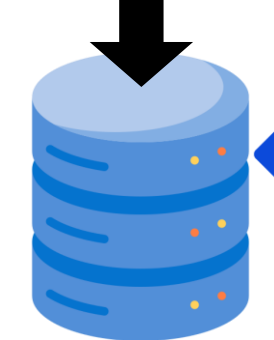
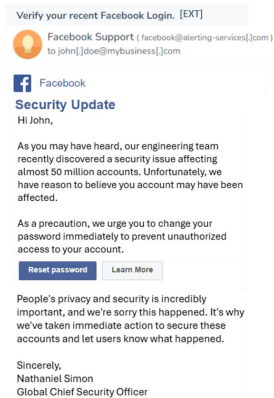
- ▶ Due to the simple-mail transfer protocol (SMTP), all emails are structured as follows:
 - ▶ **Header:**
 - ▶ **Subject:** What this email is about
 - ▶ **Sender:** Email address of the sender.
 - ▶ **Body**
 - ▶ **Body Text:** Main message of the email
 - ▶ **Signature:** May contain additional information about the sender.
 - ▶ **Attachments** (Optional): Any files included with the email
- ▶ Structured nature of emails allows features to be stored, retrieved, and used as is with minimal preprocessing needed.



Email Phishing: Data Extraction – Preprocessing (Example)

- ▶ Email databases may store emails as html/json files instead of email features, which are less structured and require file parsers for preprocessing.

Collected Raw Data (HTML)



Internal Email File Database



```
1 <!DOCTYPE html>
2 <html lang="en">
3 <head>
4   <meta charset="UTF-8">
5   <title>Security Incident 80c35a (Phishing)</title>
6 </head>
7 <body>
8   <div class="email-container">
9     <div class="header-metadata">
10      <div class="metadata-label">
11        Subject: Verify your recent
12        Facebook Login. [EXT]</div>
13      <div class="metadata-label">From:
14        Facebook Support &lt;
15        facebook@alerting-services.com></div>
16      <div class="metadata-label">To:
17        john.doe@mybusiness.com</div>
18    </div>
19    <h2 class="security-heading">Security Update
20    </h2>
21    <p>Hi John,</p>
22    <p>As you may have heard, our engineering
23    team recently discovered a security issue
24    affecting almost 50 million accounts.
25    Unfortunately, we have reason to believe
26    you account may have been affected.</p>
27    <p>As a precaution, we urge you to change
28    your password immediately to prevent
29    unauthorized access to your account.</p>
30    <p>People's privacy and security is
31    incredibly important, and we're sorry this
32    happened. It's why we've taken immediate
33    action to secure these accounts and let
34    users know what happened.</p>
35    <p>Sincerely,<br>
36    <strong>Nathaniel Simon</strong><br>
37    Global Chief Security Officer
38  </div>
39 </body>
40 </html>
```



Parser

- Common Python Tools:
- HTML parsing (bs4)
 - JSON parsing (pandas)

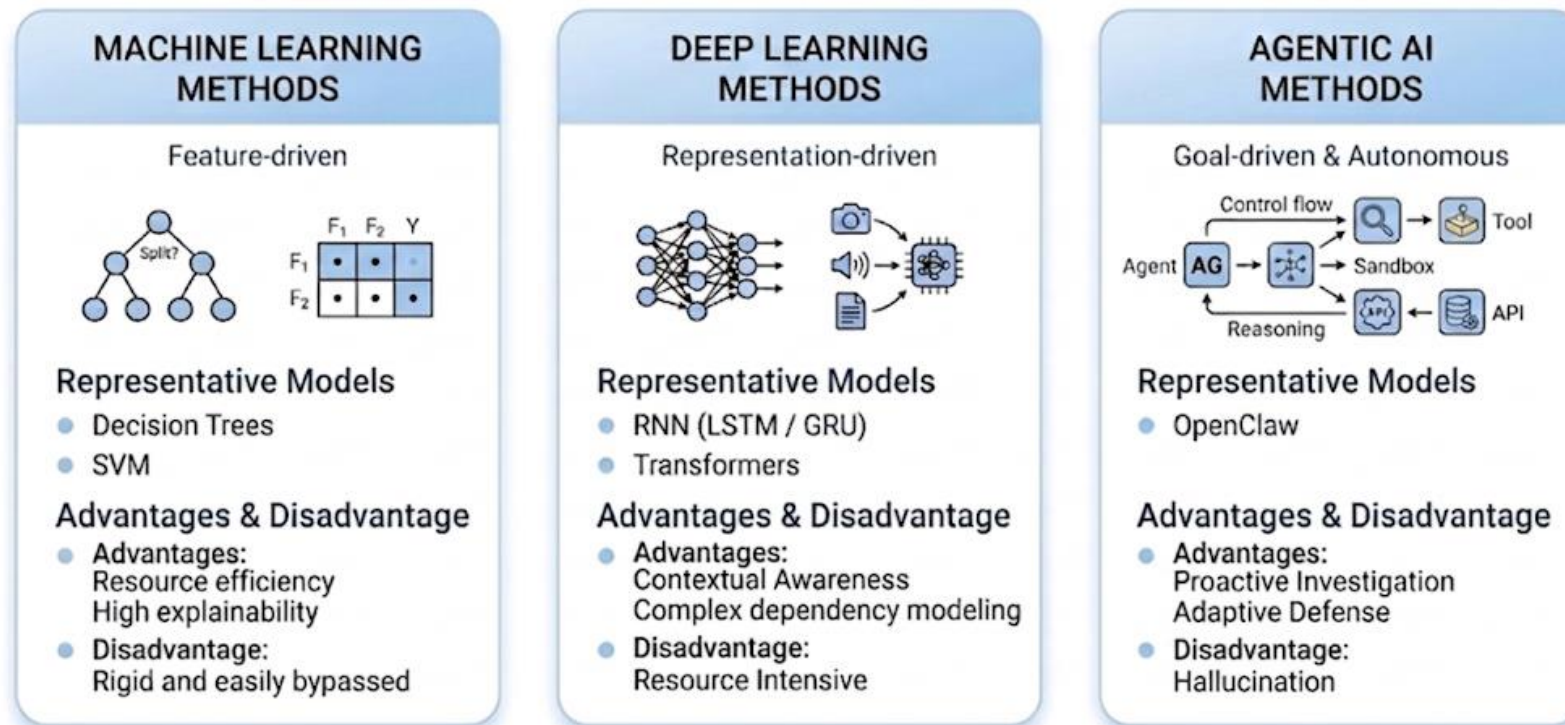
Preprocessed Datapoint

Feature	Value	Data Type
isPhish	True	Categorical
Subject	Verify your recent Facebook Login.	Textual
Sender	facebook@alerting-services.com	Textual
Body	Security Update\nHi John,\n\nAs you may have heard,...	Textual
Signature	Sincerely,\nNathaniel Simon\nGlobal Chief Security Officer	Textual
hasAtch	False	Categorical



Email Phishing: AI-Enabled Analytic

- ▶ Phishing email detection is a supervised learning task.
 - ▶ Each type of AI-enabled analytic offer different advantages and disadvantages.



■ Email Phishing: AI-Enabled Analytic – Machine Learning

- ▶ ML methods can automatically learn the pattern between numerical input features (tokenized emails) and its label (phishing or benign).
 - ▶ Low model complexity allows for fast runtimes. Certain ML models also offer some degree of explainability for their results.
- ▶ **Representative Models:**
 - ▶ Decision Tree and Support Vector Models (SVMs)
- ▶ **Challenge:** Tokenizing text data may result in loss of information.
 - ▶ **Recall:** LDAs tokenize emails into bag-of-words, which destroys the sequential dependencies in text.



Email Phishing: AI-Enabled Analytic – Machine Learning (Example Approach)

- ▶ **Idea:** Can we tokenize emails with some language meaning?
- ▶ **Approach:** Doc2Vec + PCA tokenization
 - ▶ **Doc2Vec:** An MLP that transforms textual emails into 20-D vectors.
 - ▶ **Principle Component Analysis (PCA):** Reduce dimensionality while preserving variance between datapoints (20-D $\rightarrow$ 2-D).

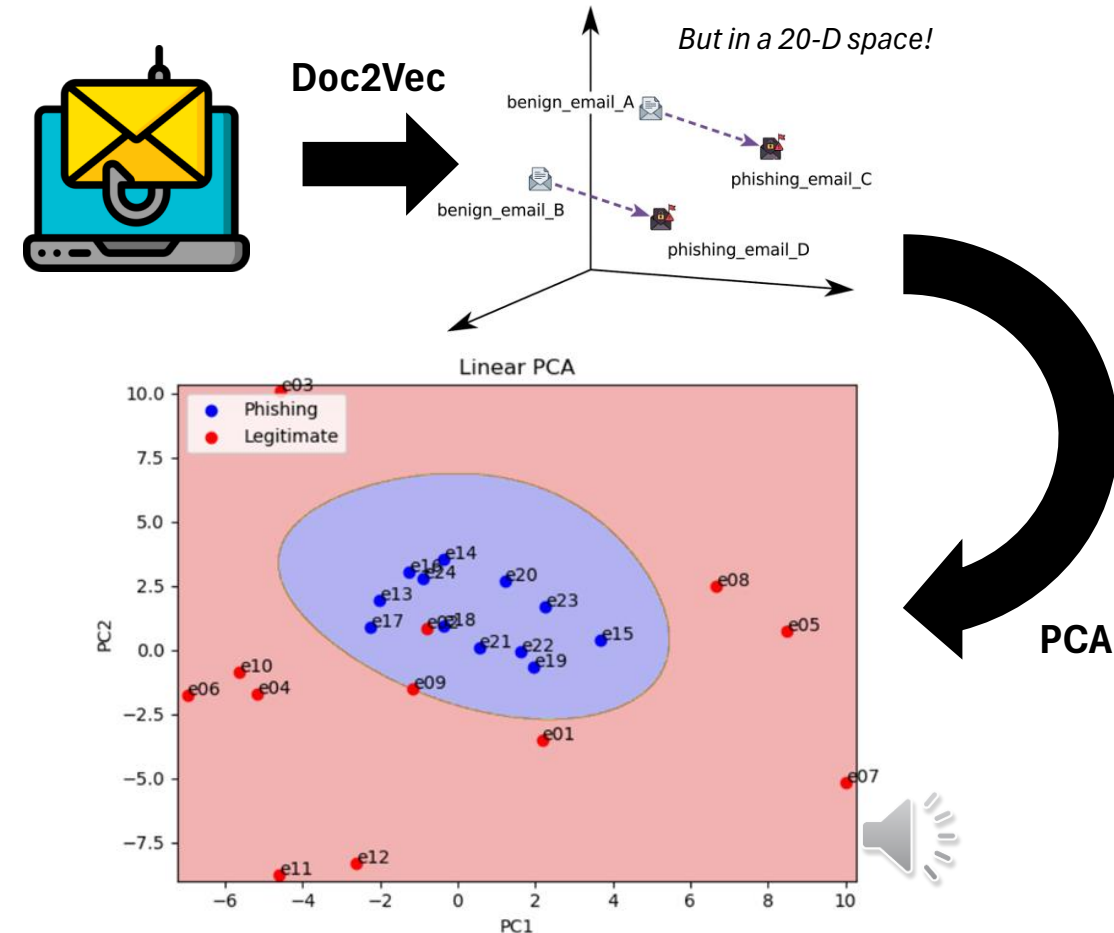


Fig. 3. SVM: decision boundary on the 2-D linear PCA projection.

Email Phishing: AI-Enabled Analytic – Machine Learning (Example Results)

- ▶ Authors trained ML classifiers on the 2-D PCA-transformed numerical input and reported the following performance metrics:
 - ▶ **Accuracy:** How many of the classifier's prediction were correct?
 - ▶ **Precision:** Out of all the emails predicted as phishing by the classifier, how many were actually phishing?
 - ▶ **Recall:** Out of all the phishing emails originally in the dataset, how many did the classifier correctly predict as phishing?
 - ▶ **F1 Score:** Balance between Precision and Recall

		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

PERFORMANCE SCORES FOR THE CLASSIFIERS IN THE 2-D LINEAR PCA SPACE.

Classifier	Accuracy	F_1 score	Precision	Recall
SVM	0.900	0.866	0.800	0.800
Logistic Regression	0.750	0.700	0.650	0.600
Random Forest	0.916	0.900	0.800	0.750
Naive Bayes	0.900	0.866	0.850	0.900



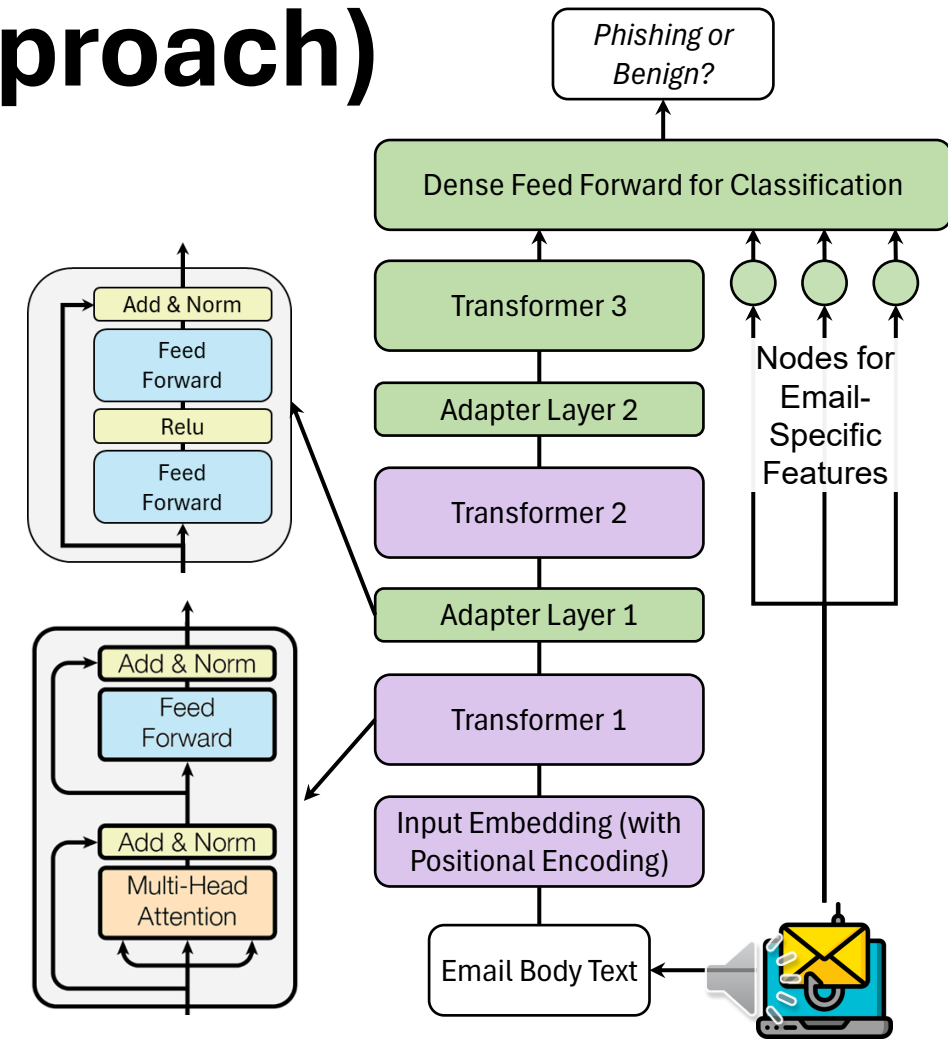
■ Email Phishing: AI-Enabled Analytic – Deep Learning

- ▶ DL models can learn sequential and long-term textual dependencies with their hidden layers and attention mechanisms
 - ▶ Does not require complicated tokenization processes.
- ▶ **Representative Models:**
 - ▶ RNNs and Transformer LLMs (BERT)
- ▶ **Challenge:** DL models can be resource intensive to train due to increased model complexity.
 - ▶ **Recall:** Training models can require up to a trillion times more GPU memory than running it.



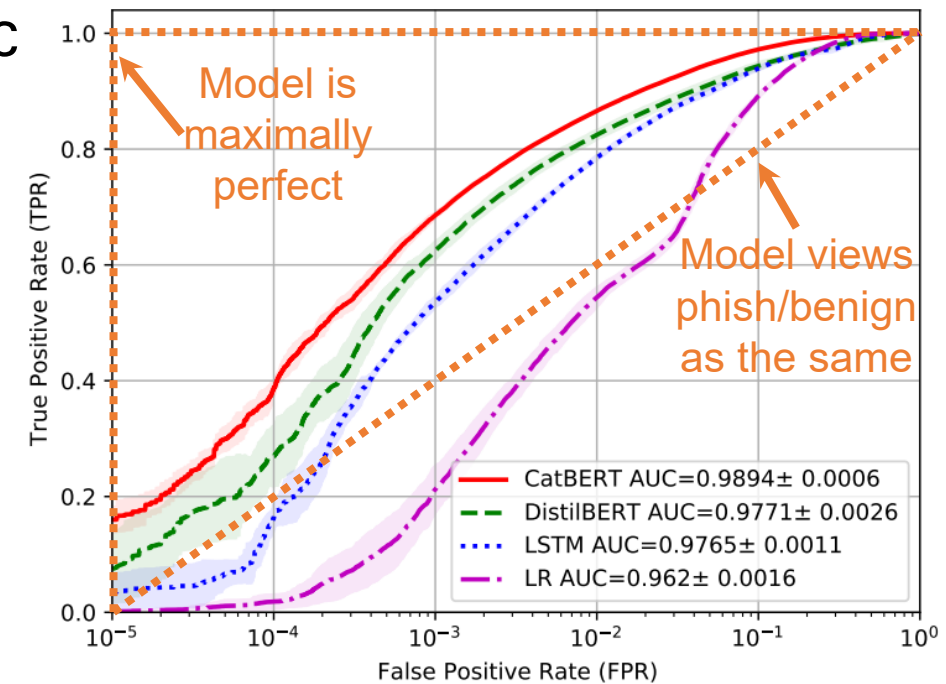
Email Phishing: AI-Enabled Analytic – Deep Learning (Example Approach)

- ▶ **Idea:** Can we just train some parts of an LLM and consider non-textual features?
- ▶ **Approach:** Context-Aware Tiny BERT (CatBERT):
 - ▶ Leverage a pretrained BERT (encoder only for text classification) and train only specific components (in green) while leaving other components frozen with pretrained parameters (in purple).
 - ▶ Add nodes to the last Feed-Forward layer for numerical email-specific features:



Email Phishing: AI-Enabled Analytic – Deep Learning (Example Results)

- ▶ Instead of numerical metrics, authors reported each method's Receiver Operating Characteristic (ROC) curves.
- ▶ **Recall:** For categorical labels, AI models output the predicted probability of each label.
 - ▶ We detokenize the model's predicted probability (numeric output) into its predicted label (human-readable output) with a threshold.
- ▶ Each point on an ROC curve is the model's performance with a different threshold.
 - ▶ Visually represent how good each model is at differentiating between phishing and benign emails.



(a) All samples

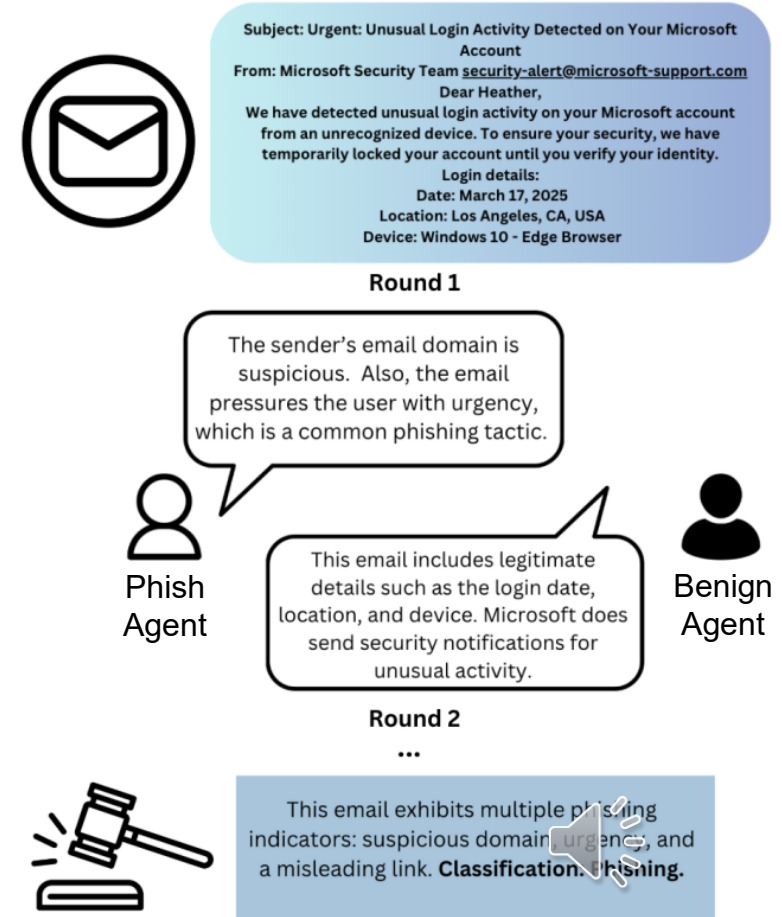


■ Email Phishing: AI-Enabled Analytic – Agentic AI

- ▶ AAI systems leverages the pretrained knowledge of LLMs to recognize phishing emails.
 - ▶ Reasoning is crucial for the agent to analyze each inputted email and adapt to adversary's new phishing techniques.
- ▶ Representative systems:
 - ▶ Single and Multi-Agent systems
- ▶ **Challenge:** AAI agents could bias its reasoning towards a decision without considering the alternative.
 - ▶ **Recall:** LLMs reason by iteratively generating thoughts and feeding those thoughts back as inputs. The bias of earlier thoughts could impact subsequent thoughts and the agent's final decision. 

Email Phishing: AI-Enabled Analytic – Agentic AI (Example Approach)

- ▶ **Idea:** Can we force the LLM to reason about both viewpoints before making its decisions?
- ▶ **Approach:** Agentic Debate for Phishing Detection
 - ▶ Inputs email into a three-agent system:
 - ▶ **Phish Agent:** Must argue that inputted email is phishing
 - ▶ **Benign Agent:** Must argue that inputted email is benign
 - ▶ **Judge Agent:** Makes the final decision on inputted email
 - ▶ Static Orchestration that models a debate with three turns:
 1. Phish and benign agents lay out their initial arguments.
 2. Phish and benign agents respond to and refute the other's argument from the 1st turn.
 3. Judge agent considers all arguments from the 1st and 2nd turns to make its decision.



Email Phishing: AI-Enabled Analytic – Agentic AI (Example Results)

- ▶ Authors leveraged different LLM-combinations for their multi-agent systems and reported performance metrics on five different phishing email datasets.

Agent Configuration			UoT		Ling		Nazario_5		Nigerian_Fraud		SpamAssassin	
Phish Agent	Benign Agent	Judge	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GPT-4	GPT-4	GPT-4	98.12%	0.98	98.76%	0.98	98.03%	0.98	98.54%	/	98.40%	0.98
LLaMA-2	LLaMA-2	LLaMA-2	98.01%	0.98	98.32%	0.98	98.22%	0.98	98.17%	/	98.09%	0.98
GPT-4	LLaMA-2	GPT-4	98.91%	0.98	99.43%	0.99	99.02%	0.99	99.27%	/	98.73%	0.98
LLaMA-2	GPT-4	GPT-4	98.36%	0.98	99.02%	0.99	98.71%	0.98	98.85%	/	99.12%	0.99
GPT-4-LLaMA-2-GPT-4-CoT			98.65%	0.98	99.12%	0.99	98.77%	0.98	99.00%	/	98.63%	0.98
GPT-4-LLaMA-2-GPT-4-Role			98.65%	0.98	98.95%	0.98	98.52%	0.98	98.74%	/	98.90%	0.99
GPT-4-LLaMA-2-GPT-4-CoT-Role			98.38%	0.98	98.77%	0.98	98.69%	0.98	98.59%	/	98.45%	0.98



Email Phishing: Deployment

- Each study leveraged holdout datasets and cross-validation to show the deployment of their AI-enabled analytic for phishing email detection.

Deployment Strategy of Example Email Phishing Detection Studies (Holdout Datasets + Cross Validation)							
Example	Reference	Original Dataset		Training Set	Validation Set	Testing Set	Cross Validation
ML Example	Gutiérrez et al., 2020	24 Hand-written Emails (Small)	12 Phishing	21-22 Datapoints (90%)	N/A	2-3 Datapoints (10%)	10-fold
			12 Benign				
DL Example	Lee et al., 2020	4.3M Internal Emails from Sophos (Large)	407K Phishing	3M Datapoints (70%)	650K Datapoints (15%)	650K Datapoints (15%)	None
			3.8M Benign				
AAI Example	Nguyen et al., 2025	17K Emails Across 5 Opensource Datasets (Medium, but diverse)	7.1K Phishing	N/A	N/A	All 17K Datapoints	None
			9.9K Benign				

Reference:

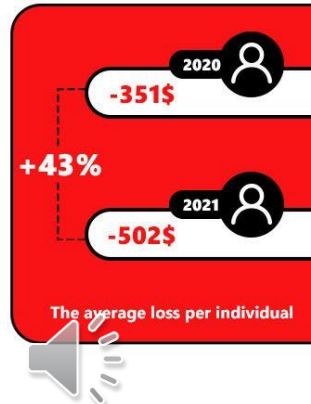
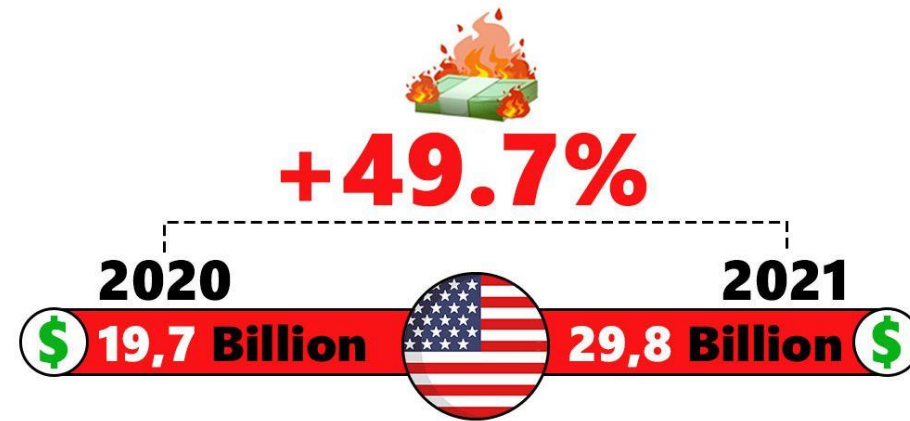
- Gutiérrez, L.F., Abri, F., Armstrong, M., Namin, A.S. and Jones, K.S., 2020, December. Email embeddings for phishing detection. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 2087-2092). IEEE.
- Lee, Y., Saxe, J. and Harang, R., 2020. CATBERT: Context-aware tiny BERT for detecting social engineering emails. arXiv preprint arXiv:2010.03484.
- Nguyen, N.T.V., Childress, F.D. and Yin, Y., 2025, April. Debate-driven multi-agent LLMs for phishing email detection. In 2025 13th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-5). IEEE.

Note: Nguyen et al. did not update any parameters. If they leveraged any APR techniques from week4's lecture, they would have needed to a training set.



Voice Phishing: Context Identification

- ▶ While phishing emails remain prevalent, other phishing modalities exist and cannot remain unaddressed.
- ▶ Voice Phishing (Vishing) is an emerging phishing attack vector.
 - ▶ Growing market for adversaries
 - ▶ Attack frequency surged by 442% in 2025 due to AI deep-fakes.
 - ▶ Global losses projected to reach \$40 billion by 2027.
 - ▶ Uniquely preys on senior citizens (40% of victims)



Americans lost 29.8 billion USD to vishing in 2021, a 49.7% increase from the 19.7 billion USD in 2020.
The average loss per individual was 502 USD in 2021, up 43% from 2020.

Voice Phishing: Context Identification

- ▶ Vishing detection presents unique challenges:
 - ▶ **Real-time Detection:** Victims are spurred into action mid-conversation and not after receiving an email. Detection methods cannot analyze the entire conversation before it happens!
 - ▶ **Non-Verbal Persuasion:** Adversaries can leverage vocal techniques outside of transcript content (pitch, tone, etc.) to convey/amplify persuasion cues to the victim.

1. Reciprocity

Persuader makes a 'gift' like offer of free assistance or a special benefit.



2. Commitment

Persuader uses small initial agreements ('yes' responses) to lock the victim into larger follow-up actions.



3. Social Proof

Persuader mentions others, like coworkers or similar businesses, who have already complied.



4. Liking

Persuader uses compliments, commonalities, or perceived similarity to build rapid, manipulative rapport.



5. Authority

Persuader adopts the voice and visual cues of a specific, high-stakes role (e.g., bank fraud investigator, law enforcement).



6. Scarcity

Persuader presents an artificial 'urgent limited-time offer' or 'one-time penalty' that will expire shortly.



Voice Phishing: Context Identification

- Current vishing detection approaches typically fall under two categories:



MANUAL TRAINING & EDUCATION

APPROACH:

- Employee workshops,
- phishing simulations, and awareness campaigns



MAJOR LIMITATIONS:

- 1. CANNOT BE SCALED:**
Resource-intensive for large organizations.
- 2. HUMANS ARE UNRELIABLE DETECTORS:**
Prone to manipulation, fatigue, and emotional triggers.



TEXT CONVERSION + CLASSIFICATION

APPROACH:

- Speech-to-text transcription,
- automated text analysis, and pattern matching



MAJOR LIMITATIONS:

- 1. MISSES NON-VERBAL FEATURES:**
Fails to analyze pitch, tone, pace, and other acoustic nuances.





Voice Phishing: Context Identification



Stakeholder

- **Primary User:** Individuals and employees targeted by vishing attacks (e.g., senior citizens).
- **Critical Asset:** Sensitive and private information.



Task: Vishing Detection

- **Objective:** Real-time vishing detection before the victim is tricked into information disclosure
- **Current Data:** Phishing email datasets.



Current Approaches

- **Core Question:** Is this current conversation a vishing attack?
- **Current Method:** Individual education. Textual conversion and classification



Current Limitations

- **Data-Wise:** Large vishing datasets are not widely available and do not include deep-fakes
- **Method-Wise:** Individuals are inconsistent. Textual conversion misses prosodic features

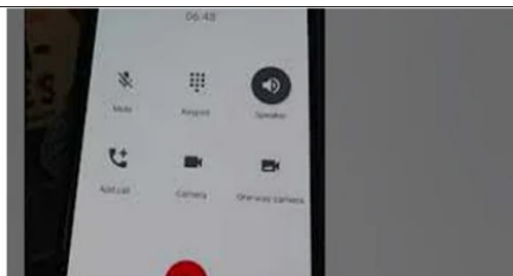


Voice Phishing: Data Extraction – Collection

- ▶ Unlike phishing emails, vishing data does not exist in curated datasets and need to be gathered via open-source intelligence (OSINT).
- ▶ Common sources (OSINT):
 - ▶ Social Media Platforms
 - ▶ YouTube Videos and Twitch VODs (Call Audio)
 - ▶ Forum Posts (Textual Transcripts)
 - ▶ Public RoboCall Datasets
- ▶ Collection methods: Web and API crawling



What an IRS Imposter Scam Sounds Like
38K views • 7 years ago



Bank of America scam call#1
3.9K views • 4 years ago

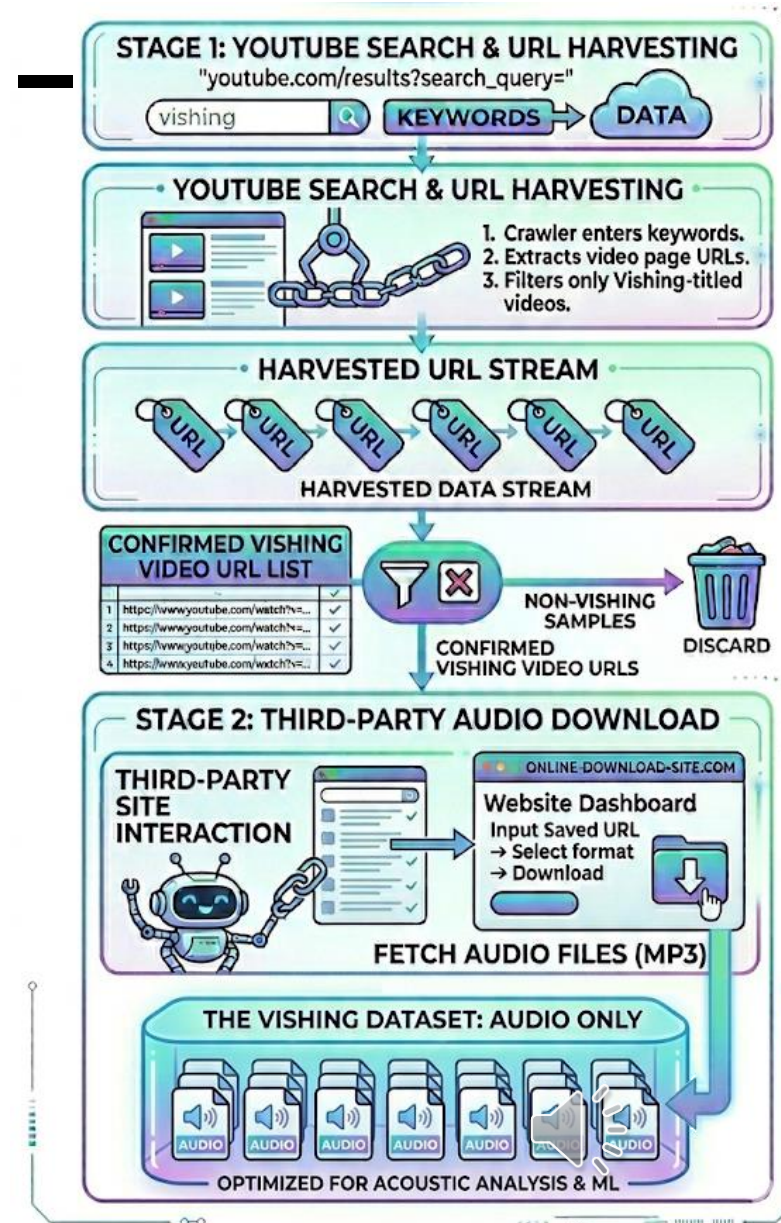


IT'S A SCAM! MEDICARE SCAM!
3.7K views • 5 months ago



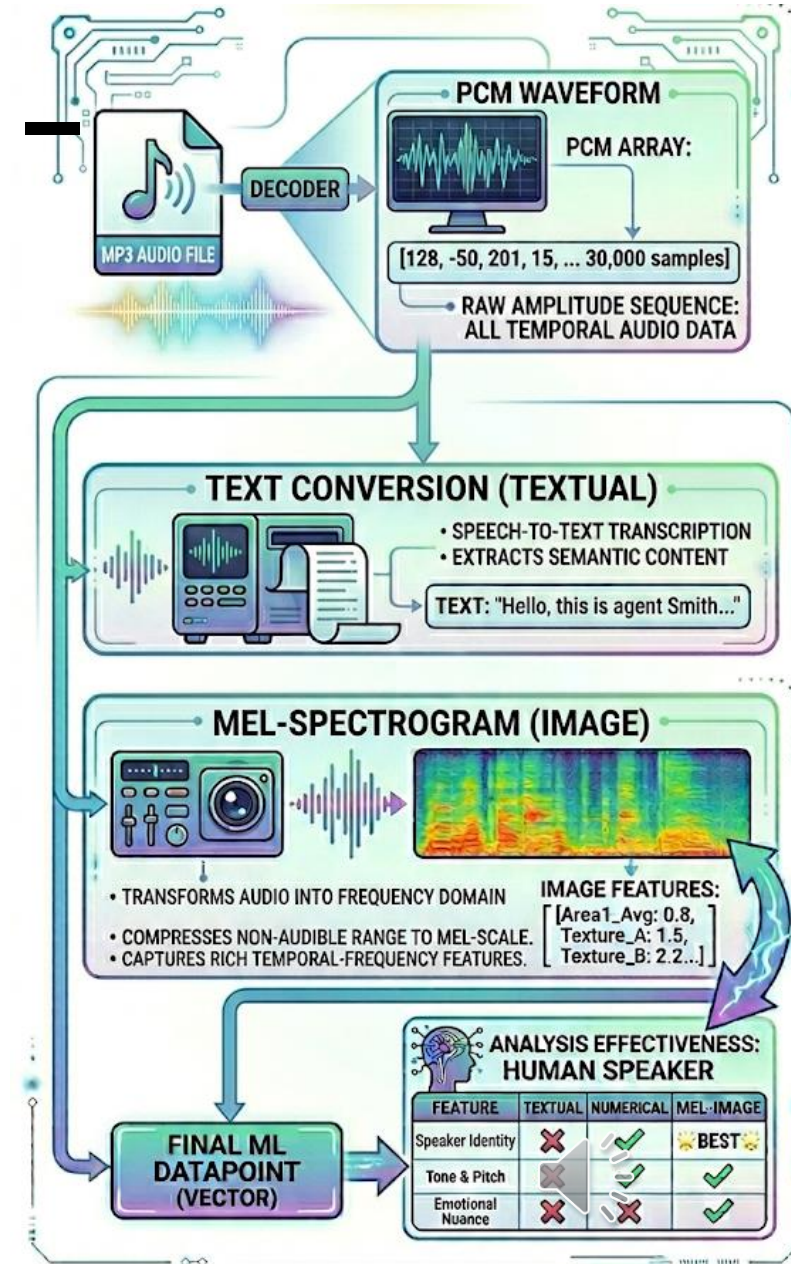
Voice Phishing: Data Extraction – Collection (Example)

- ▶ YouTube is a good OSINT source for vishing.
 - ▶ However, YouTube also contains unrelated data and requires targeted crawling.
- ▶ Keywords can constrain the YouTube search to mostly vishing videos.
 - ▶ Manual confirmation can filter out the remaining non-vishing samples.
 - ▶ Third-party websites can download audio files with the crawled links.



Voice Phishing: Data Extraction – Preprocessing

- ▶ Raw audio is stored as Pulse-Code Modulation (PCM) Waveforms and often contain too much data for direct usage.
 - ▶ PCM Waveform [*Numerical*]: Vibration measured by the microphone.
 - ▶ **Consider:** Human hearing only accounts for a certain frequency range (20 to 20K Hertz). Any data outside of that range is inconsequential for vishing detection.
- ▶ Audio preprocessing techniques include:
 - ▶ Text Conversion [*Textual*]: Focus on audio transcripts
 - ▶ Mel-Spectrogram [*Image*]: Heatmap of different human-hearable frequencies across time.



Mel-Spectrograms is the best technique for human speaker analysis by capturing temporal-frequency information



Voice Phishing: AI-Enabled Analytic

- ▶ Vishing detection is also a supervised learning task. However, unique constraints limit the type of appropriate AI-enabled analytic:
 - ▶ ML models struggle learning complex patterns between a mel-spectrogram (preprocessed audio datapoint) and a label (vishing/benign) due to model simplicity.
 - ▶ AAI excel at analyzing textual transcripts, but will miss the non-verbal cues present in mel-spectrograms



Voice Phishing: AI-Enabled Analytic – Deep Learning

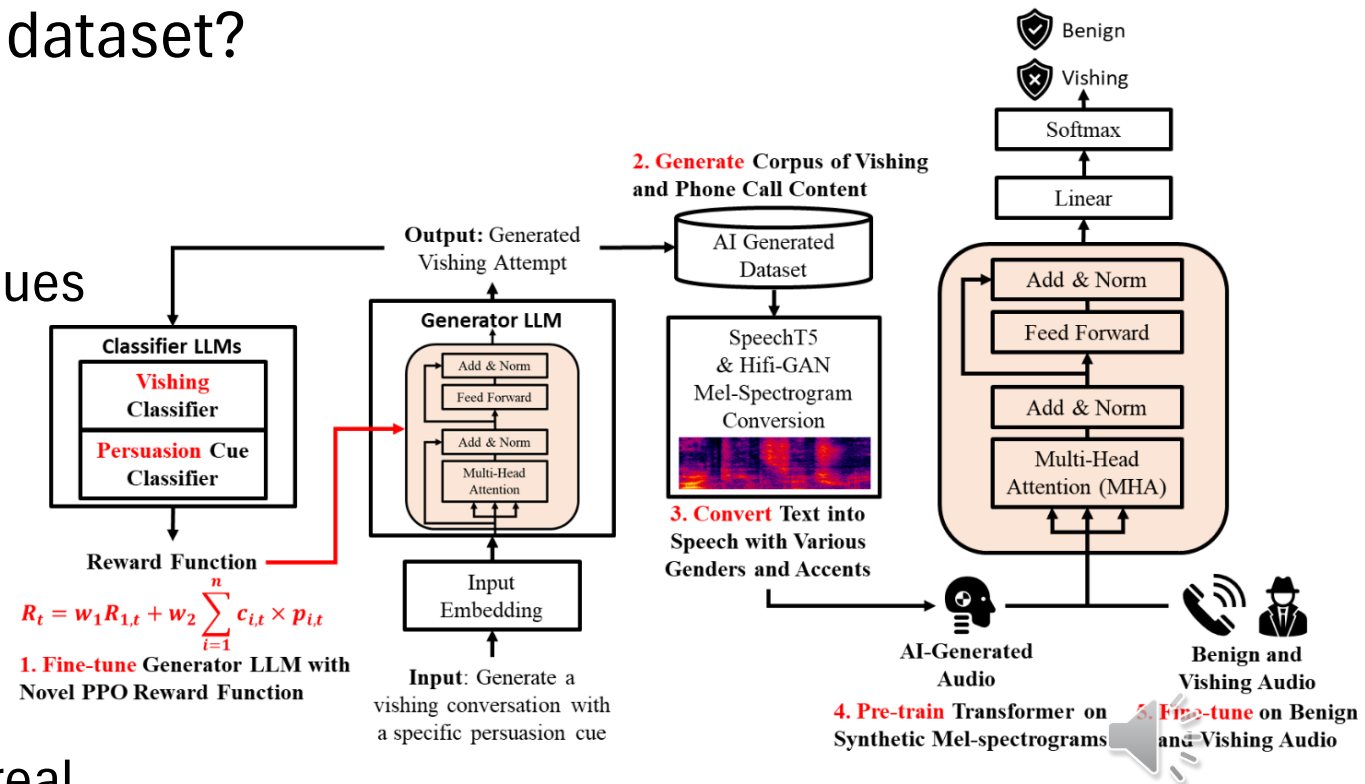
- ▶ DL models are the current most suitable methods due to being able to learn specific tasks through task-adaptation.
- ▶ **Representative Models:**
 - ▶ RNNs and Transformer LLMs (BERT)
- ▶ **Challenge:** Training DL models require a significant volume of data.
 - ▶ **Recall:** Vishing data must be collected and manually validated. Scaling vishing data extraction for an adequately-large training dataset is unfeasible.
 - ▶ Additionally, adversaries are now using deepfake models to generate vishing attacks, which may not be reflected in vishing sources.

Voice Phishing: AI-Enabled Analytic – Deep Learning (Example Approach)

- ▶ **Idea:** If adversaries are using deepfakes models to generate vishing attacks, why not do the same for a training dataset?

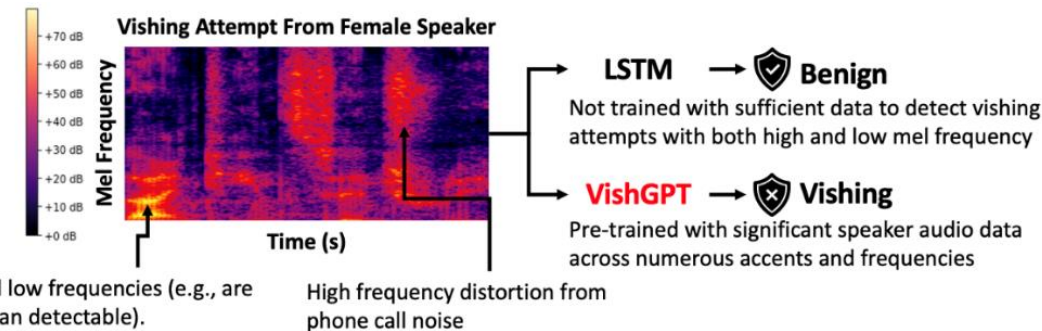
- ▶ **Approach:** VishGPT

1. Finetune LLM to generate vishing transcripts based on persuasion cues
2. Generate dataset of vishing transcripts
3. Convert transcripts to mel-spectrograms
4. Pretrain a transformer model on synthetic dataset
5. Finetune transformer model with real vishing data

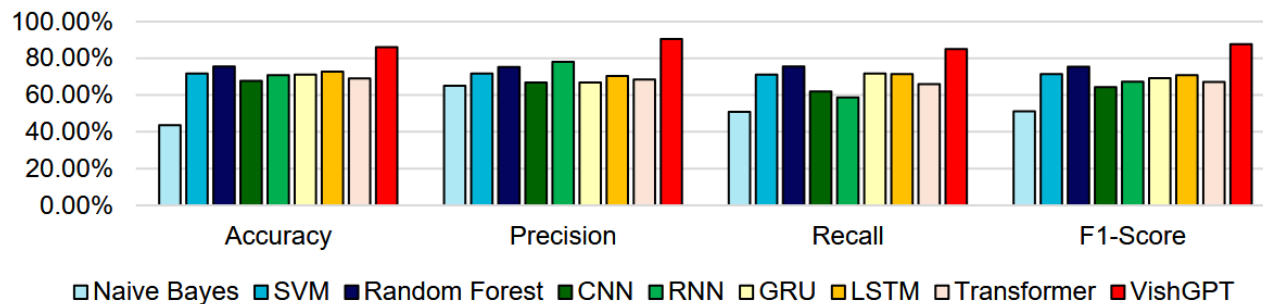


Voice Phishing: AI-Enabled Analytic – Deep Learning (Example Results)

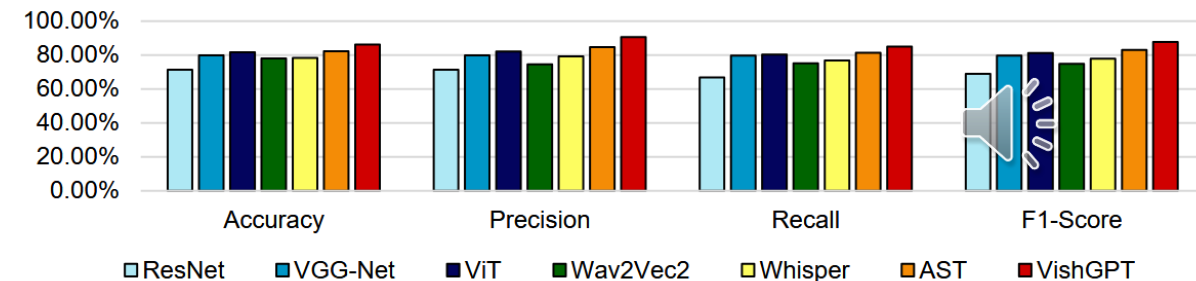
- ▶ VishGPT is a transformer model with a unique pretraining procedure.
- ▶ Authors showed how both aspects contributed to VishGPT's state-of-the-art vishing detection performance:
 - ▶ Experiment 1: Effect of the Transformer model.
 - ▶ Experiment 2: Effect of the pretraining procedure.



Experiment 1: VishGPT Against Benchmark ML and DL Models



Experiment 2: VishGPT Against Pre-Trained Vision and Audio Models



Voice Phishing: Deployment

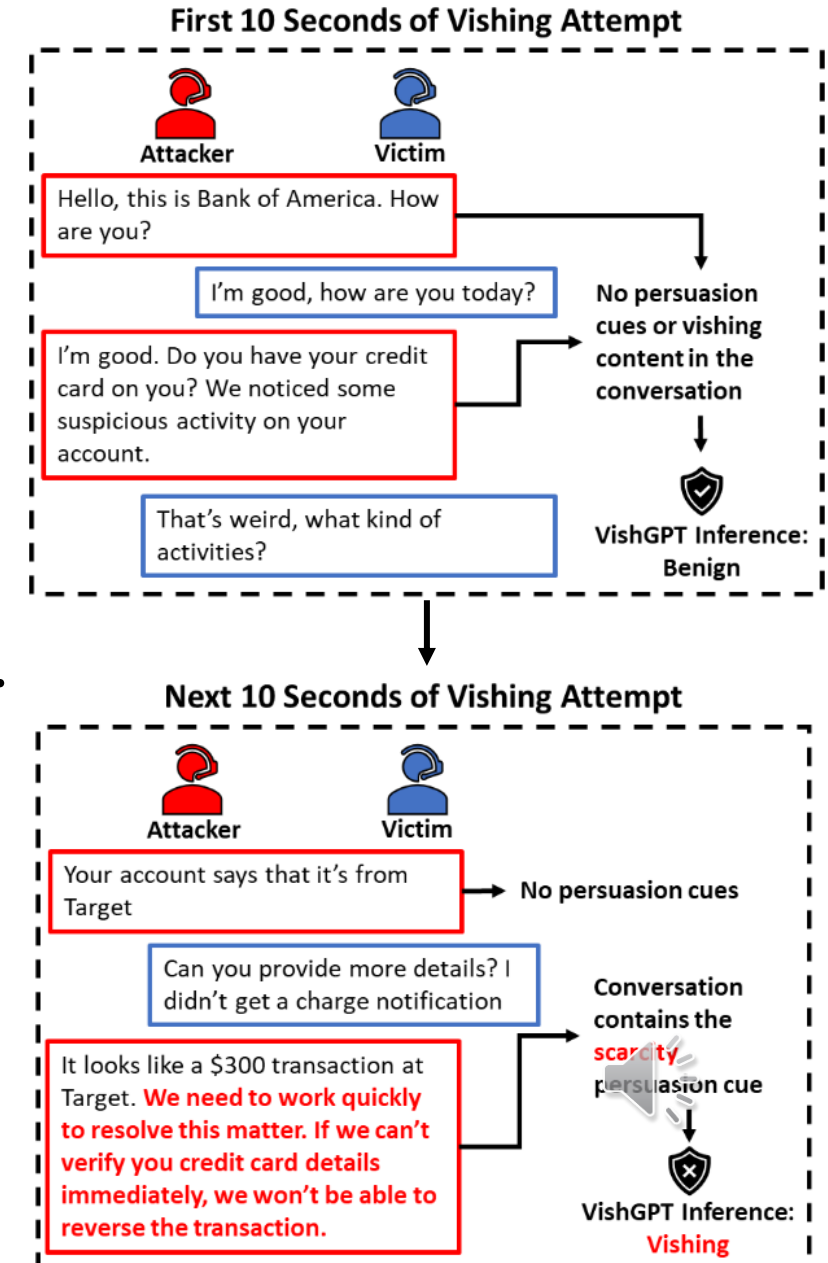
- Similar to the email phishing detection examples, authors leveraged holdout datasets to demonstrate VishGPT's deployment performance:

Deployment Strategy of Example Vishing Detection Study (Holdout Datasets + Cross Validation)						
Reference	Original Dataset		Training Set	Validation Set	Testing Set	Cross Validation
Ampel et al., 2026	3.6M Synthetic Mel-Spectrograms (Large)	1.6M Vishing	All 3.6M Datapoints	N/A	N/A	N/A
		1.6M Benign				
	3,559 YouTube Mel-Spectrograms (Small)	1,251 Vishing	3203 Datapoints (90%)	N/A	356 Datapoints (10%)	10-fold
		2,308 Benign				

Note: Pretraining does not require validation and testing sets as the goal is to learn a general pattern and not specific task performance.

Voice Phishing: Deployment

- ▶ Authors also constructed a situated implementation case study.
- ▶ **Consider:** Senior citizens (VishGPT's key stakeholders) would download VishGPT as an app that monitors their phone calls (real-time detection).
 - ▶ However, VishGPT's development environment were mel-spectrograms of whole conversations.
- ▶ Authors implemented sliding windows to simulate VishGPT analyzing parts of a phone call in real-time.
 - ▶ VishGPT still shows good performance even in the simulated case study environment!



Review Questions

- ▶ What are holdout datasets? What primarily determines between using validation sets or cross validation?
- ▶ What are the two different types of case studies? Compare the advantages and disadvantages between the two case study types.
- ▶ Why is the phishing context important?
 - ▶ Do you have a similar motivation in your projects?
- ▶ Why did vishing detection require specific types of AI-enabled analytics?
 - ▶ Do you have a similar justification for method choice in your projects?



Next Week

