Benchmarking Large Language Models for Zero-shot and Few-shot Phishing URL Detection

Najmul Hasan, Prashanth BusiReddyGari

Paper

Problem

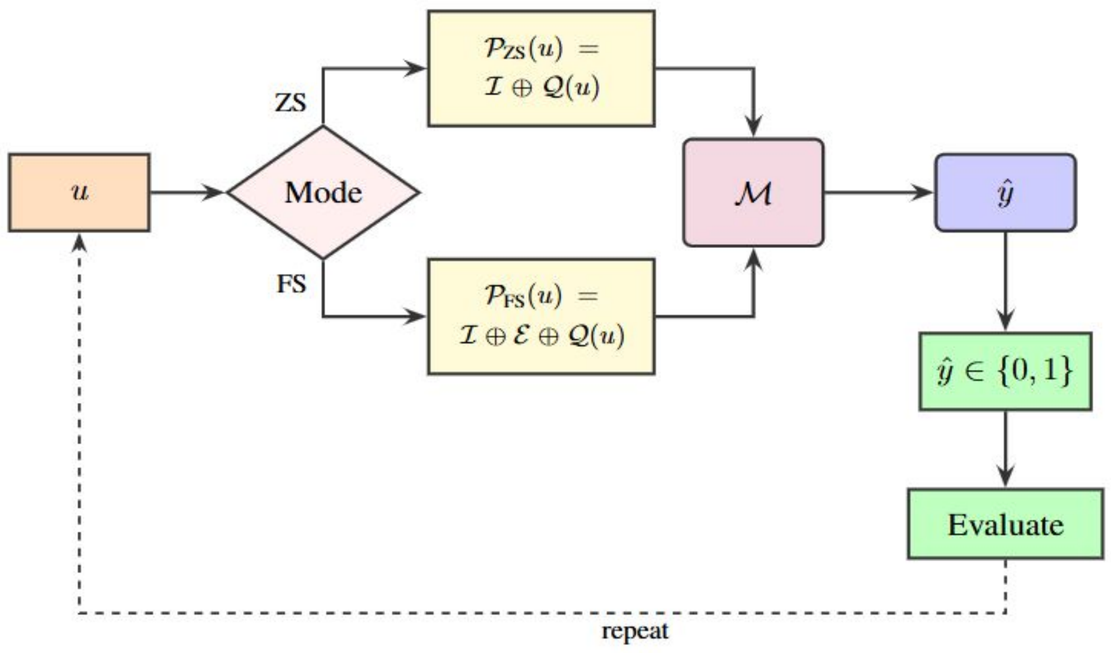
- Phishing volume up 4,000% since 2022
- ~50% of attacks evading traditional detection
- AI-generated phishing sites virtually indistinguishable
- Blacklist methods fail on newly generated URLs
- No unified LLM benchmark exists for this task

Experimental Setup

- Dataset: PhiUSIIL Phishing URL
- Balanced: 10,000 URLs (50% phishing)
- Imbalanced: 1,000 URLs (1% & 10% phishing)
- Models: GPT-4o, Claude-3.7-sonnet, Grok-3-Beta
- Few-shot: 6 examples (3 phishing + 3 legitimate)
- Metrics: Accuracy, Precision, Recall, F1, AUROC, AUPRC

Phishing URL classification methodology

Phishing URL classification pipeline under zero-shot (P_ZS) and few-shot (P_FS) prompting



Prompt Construction

ZERO-SHOT

I: "You are a cybersecurity expert. Respond only with 0 for phishing or 1 for legitimate."

Q(u):
"URL: {u}
Is this URL phishing or legitimate?
Respond with 0 or 1."

$\hat{y} \in \{0, 1\}$

$P\_ZS(u) = I \oplus Q(u)$

FEW-SHOT

I: "You are a cybersecurity expert. Respond only with 0 for phishing or 1 for legitimate."

E: 6 examples
(3 phishing + 3 legitimate)

Format:
"URL: {u'}
Answer: {y'}"

Q(u):
"URL: {u}..."

$\hat{y} \in \{0, 1\}$

$P\_FS(u) = I \oplus E \oplus Q(u)$

where ⊕ denotes concatenation

Key Findings

Balanced Dataset (10,000 URLs):

- Few-shot (6 examples) improves ALL models
- Grok-3-Beta: Best accuracy (0.9405) & F1 (0.9399)
- Claude-3.7: Best recall (0.9526)
- False negatives reduced: Grok-3 drops 950 to 248

Imbalanced Dataset (1% & 10% phishing):

- Models robust under extreme class imbalance
- Grok-3-Beta zero-shot: 0.976 accuracy, 0.938 recall
- Few-shot consistently improves F1 scores

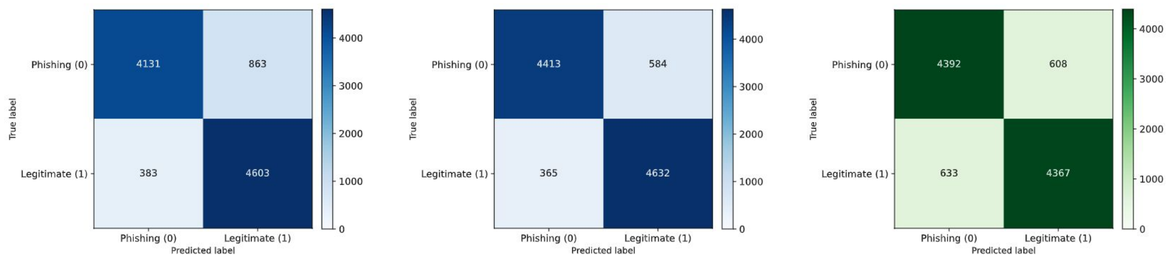
Experimental Results

Performance comparison on balanced phishing URL detection (10,000 URLs)

Model	Setting	Accuracy	Precision	Recall	F1	AUROC	AUPRC
GPT-4o	Zero-shot	0.8752	0.8421	0.9232	0.8808	0.8752	0.9018
GPT-4o	Few-shot	0.9050	0.8880	0.9270	0.9071	0.9050	0.9258
Claude-3.7	Zero-shot	0.8759	0.8778	0.8734	0.8756	0.8759	0.9072
Claude-3.7	Few-shot	0.9250	0.9027	0.9526	0.9270	0.9250	0.9395
Grok-3-Beta	Zero-shot	0.8914	0.8361	**0.9735**	0.8996	0.8914	0.9114
Grok-3-Beta	Few-shot	**0.9405**	**0.9492**	0.9307	**0.9399**	**0.9405**	**0.9573**

Per-Class Evaluation

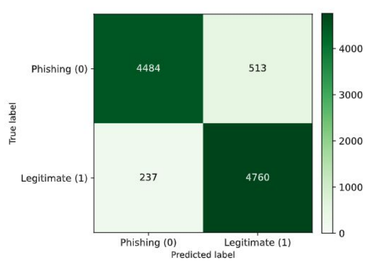
Confusion matrices for all models under zero-shot and few-shot prompting



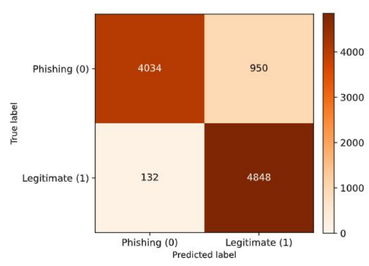
(a) GPT-4o: Zero-shot

(b) GPT-4o: Few-shot

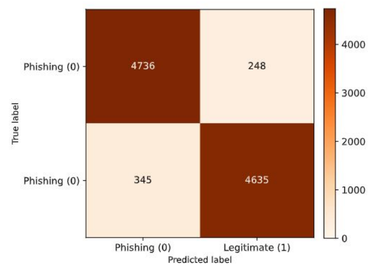
(c) Claude: Zero-shot



(d) Claude: Few-shot



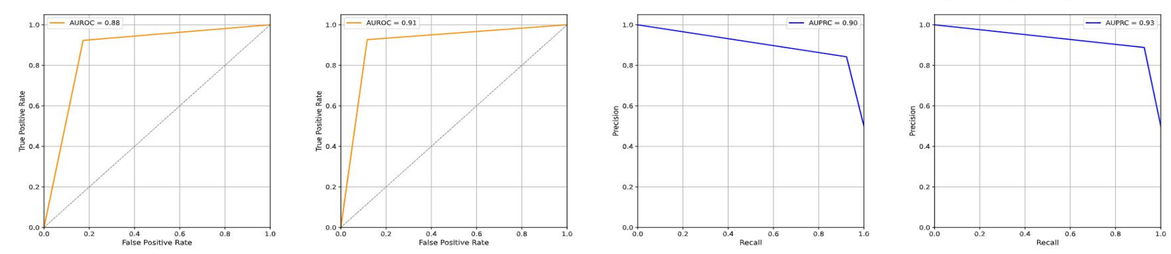
(e) Grok-3: Zero-shot



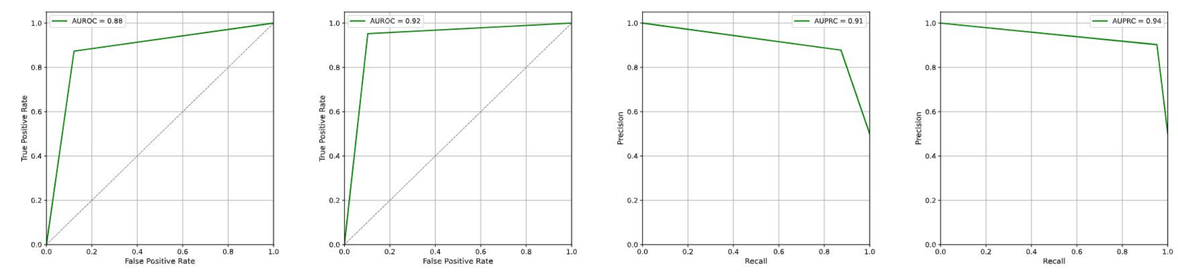
(f) Grok-3: Few-shot

AUROC and AUPRC

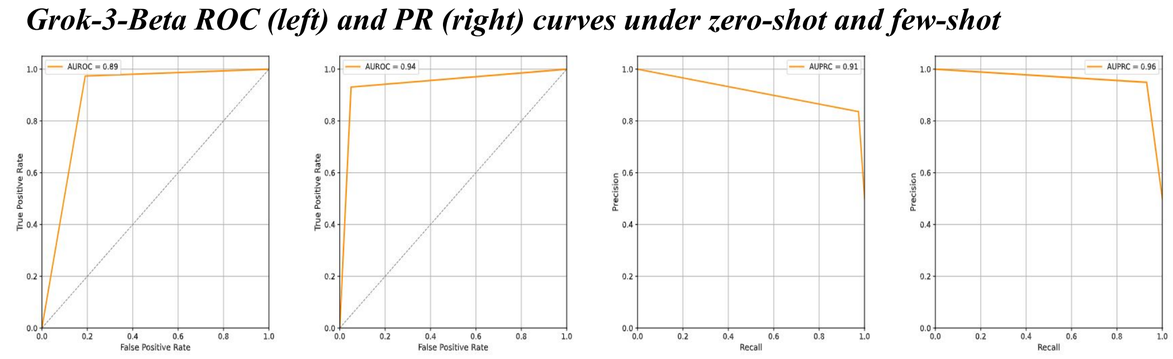
GPT-4o ROC (left) and PR (right) curves under zero-shot and few-shot prompting.



Claude-3.7 ROC (left) and PR (right) curves under zero-shot and few-shot prompting.



Grok-3-Beta ROC (left) and PR (right) curves under zero-shot and few-shot



Imbalanced Dataset Results

Model	Metric	Zero-Shot				Few-Shot		
		S123-1%	S123-10%	S456-1%	S456-10%	Ɛ=1	Ɛ=3	Ɛ=9
GPT-4o	Accuracy	0.917	0.902	0.935	0.927	0.833	0.908	0.942
	Precision	0.544	0.742	0.561	0.789	0.676	0.754	0.821
	Recall	0.859	0.866	0.918	0.888	0.863	0.891	0.919
	F1-Score	0.559	0.785	0.591	0.828	0.709	0.801	0.861
Claude-3.7	Accuracy	0.881	0.879	0.903	0.903	0.945	0.933	0.951
	Precision	0.535	0.716	0.542	0.749	0.837	0.801	0.846
	Recall	0.890	0.879	0.902	0.911	0.881	0.905	0.915
	F1-Score	0.534	0.761	0.553	0.799	0.857	0.842	0.876
Grok-3-Beta	Accuracy	0.964	0.945	0.976	0.962	0.969	0.915	0.924
	Precision	0.602	0.839	0.640	0.881	0.949	0.768	0.783
	Recall	0.932	0.872	0.938	0.921	0.872	0.931	0.918
	F1-Score	0.657	0.854	0.708	0.900	0.906	0.821	0.831