← Back to **Author Console** (/group?id=NeurIPS.cc/2025/Conference/Authors#your-submissions)

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

(/pdf?id=pkJT7Vegrz)

Najmul Hasan (/profile?id=~Najmul_Hasan1),
Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1)

11 May 2025 (modified: 18 Sept 2025) Submitted to NeurIPS 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=pkJT7Vegrz) BibTeX

Keywords: AI in Cybersecurity (AICyber), Phishing Detection, Large Language Models, Zero-shot Learning, Few-shot Learning, Prompt Engineering

TL;DR: Zero-shot and few-shot learning with LLMs offers practical utility, operational efficiency, and scalability. We benchmarked three LLMs for phishing URL detection using zero-shot and few-shot prompting.

Abstract:
The Uniform Resource Locator (URL), introduced in a connectivity-first era to define access and locate resources, remains historically limited, lacking future-proof mechanisms for security, trust, or resilience against fraud and abuse, despite the introduction of reactive protections like HTTPS during the cybersecurity era. In the current AI-first threatscape, deceptive URLs have reached unprecedented sophistication due to the widespread use of generative AI by cybercriminals and the AI-vs-AI arms race to produce context-aware phishing websites and URLs that are virtually indistinguishable to both users and traditional detection tools. Although AI-generated phishing accounted for a small fraction of filter-bypassing attacks in 2024, phishing volume has escalated over 4,000% since 2022, with nearly 50% more attacks evading detection. At the rate the threatscape is escalating, and phishing tactics are emerging faster than labeled data can be produced, zero-shot and few-shot learning with large language models (LLMs) offers a timely and adaptable solution, enabling generalization with minimal supervision. Given the critical importance of phishing URL detection in large-scale cybersecurity defense systems, we present a comprehensive benchmark of LLMs under a unified zero-shot and few-shot prompting framework and reveal operational trade-offs. Our evaluation uses a balanced dataset with consistent prompts, offering detailed analysis of performance, generalization, and model efficacy, quantified by accuracy, precision, recall, F1 score, AUROC, and AUPRC, to reflect both classification quality and practical utility in threat detection settings. We conclude few-shot prompting improves performance across multiple LLMs.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Financial Support: nh0033@bravemail.uncp.edu
Reviewer Nomination: prashanth.busireddygari@uncp.edu, nh0033@bravemail.uncp.edu
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Applications (e.g., vision, language, speech and audio, Creative AI)
LLM Usage: Editing (e.g., grammar, spelling, word choice)
Declaration: I confirm that the above information is accurate.
Submission Number: 23634

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
Everyone Program Chairs Submission23634... Submission23634... Submission23634... Submission23634... Submission23634... Submission23634... Submission23634... Submission23634...
29 / 29 replies shown
Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 17 Sept 2025, 08:52 (modified: 18 Sept 2025, 10:33) Program Chairs, Authors
Revisions (/revisions?id=G1WJPszMp1)

Decision: Reject

Comment:
The paper carries out an evaluation of the effectiveness of large-language models (LLMs) for detecting malicious URL. Detection of malicious URLs is a long-standing problem, and such a domain is in desperate need of solutions that work (i.e., can detect malicious URLs without triggering false positives).

Four reviewers provided their recommendations and also interacted with the authors. While one reviewer was positive, three expressed substantially-negative recommendations (which did not change despite the interaction with the authors). In particular, the most noteworthy cause of concern is the lack of comparison with baselines: the paper merely "benchmarks LLMs", which is not bad per-se, but it is when considering the plethora of work (just perform a Google-Scholar search with the query "malicious URL detection") that has addressed this problem. Therefore, the results show that some LLMs/techniques are better than others---but are such techniques better than what is already available? Lack of a comprehensive assessment on a standardized testbed prevents determining how significant the findings of this paper can be.

Due to the above, I recommend rejection of this paper. However, I believe that the reviews were overall of very good quality and I encourage the authors to account for the reviewers' feedback in improving their research. On this note, I also suggest to perhaps consider a different venue (e.g., a security-focused venue may be more appropriate to communicate new results of empirical nature---if such findings have practical relevance!).

Author Final Remarks by Authors

Author Final Remarks by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))
11 Aug 2025, 10:19 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Author Final Remarks:

The novelty of our work is in its empirical and operational focus, consistent with other benchmarking studies published at NeurIPS and related venues. Benchmarking can be as impactful as methodological innovation when it establishes a realistic evaluation setting and yields insights unavailable from prior studies.

This study presents one of the first reproducible benchmarks of proprietary, closed-weight LLMs for phishing URL detection within constraints representative of real-world deployment. The evaluation is conducted entirely via public APIs, without weight access or fine-tuning, using a standardized prompt and metric setup to ensure consistent comparison across heterogeneous API schemas. It incorporates robustness testing under class imbalance and latency/cost profiling, producing metrics relevant to production.

Our aim was to address a gap not covered in Nasution et al. or other prior work. No controlled, large-scale, operational benchmark exists for leading proprietary LLM APIs on phishing URL detection under deployment conditions that include heterogeneous API schemas, strict latency budgets, rate limits, and no ability to fine-tune. These are not incidental variations; they fundamentally shape both methodology and the nature of the insights obtained.

The contribution is in the evaluation framework and the empirical evidence it produces. Just as benchmarks such as GLUE (Wang et al., 2018) or MMLU (Hendrycks et al., 2021) were impactful not because they introduced new model architectures but because they revealed performance differences under standardized testing, this work offers new knowledge to researchers and practitioners in cybersecurity. It provides empirical answers to a question the community has not previously had the basis to address, offering evidence absent from both the literature and practitioner resources.

The work is relevant to NeurIPS because it addresses the methodological challenge of evaluating API-only, closed-weight LLMs in a security domain, an increasingly important scenario for both research and deployment. Our standardized framework ensures identical task semantics while adapting to API constraints, enabling consistent cross-model comparison. By integrating robustness testing under class imbalance and latency/cost profiling, the study yields deployment-relevant metrics absent from prior literature.

The manuscript was prepared using the official NeurIPS submission template, adhering to all formatting requirements.

Official Review of Submission23634 by Reviewer D1vP

Official Review by Reviewer D1vP 02 Jul 2025, 13:06 (modified: 18 Sept 2025, 13:00)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer D1vP Revisions (/revisions?id=Vb8fkk4uoc)

Summary:
The authors present a comprehensive benchmark evaluation of proprietary large language models (LLMs) for phishing URL detection under zero-shot and few-shot learning settings. The authors systematically compare three commercial LLMs (GPT-4o, Claude-3.7-sonnet, and Grok-3-Beta) using a balanced dataset of 10,000 URLs with standardized prompts and evaluation metrics. Key findings demonstrate that few-shot prompting with just six examples significantly improves detection performance across all models, with Grok-3-Beta achieving the highest accuracy (0.9405) and Claude-3.7-sonnet showing the best recall (0.9526). The work provides valuable insights into the practical utility of prompt-based LLMs for cybersecurity applications.

Strengths And Weaknesses:
Strengths

1. The paper addresses a timely and important problem in cybersecurity, where AI-generated phishing attacks are becoming increasingly sophisticated and difficult to detect using traditional methods.
2. The experimental design is rigorous, with careful attention to dataset balancing, prompt standardization, and comprehensive evaluation metrics (including AUROC and AUPRC).
3. The comparative analysis reveals interesting model-specific trade-offs (e.g., Grok-3-Beta's precision vs Claude-3.7-sonnet's recall) that could guide practical deployment decisions.
4. The methodology section is particularly thorough, detailing prompt construction, API protocols, and evaluation procedures to support reproducibility.

Weakness

1. The evaluation uses a balanced dataset (50% phishing URLs), which doesn't reflect the extreme class imbalance typically found in real-world scenarios (where phishing URLs are rare).
2. The paper doesn't discuss computational costs or latency considerations, which are critical for real-time phishing detection systems.
3. Limited analysis is provided about why certain models perform better on specific metrics (e.g., what makes Claude better at recall).
4. The fixed k=6 few-shot examples may not be optimal - some exploration of different k values could strengthen the findings.

Quality: 3: good
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:
1. Could you evaluate performance on an imbalanced test set to better reflect real-world conditions?
2. Have you considered the inference latency and computational costs of using these commercial LLMs at scale?
3. What architectural or training differences might explain the observed model-specific tradeoffs?
4. Did you experiment with different numbers of few-shot examples (k) to determine optimal context size?
5. Could you discuss potential adversarial attacks that might bypass your detection method?

Limitations:
Yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
N/A

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
Although the authors address my concerns, I still maintain my score due to the large latency, which hinders the practicality of the proposed method in the real-world scene.

Rebuttal by Authors

Rebuttal by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

29 Jul 2025, 18:41 (modified: 31 Jul 2025, 16:54) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=EfyWDNgsEa)

Rebuttal:

Thank you for your constructive review.

W1. Our choice to begin with a balanced dataset was motivated to control for class distribution effects and focus on evaluating the intrinsic classification capability of our model. Balanced testing ensures models are not reliant on majority-class bias and gives a clearer picture of class-separation ability. Such a setup is commonly used in the literature (Hannousse & Yahiouche, 2021) and facilitates fair benchmarking across methods without bias from prior probabilities. We recognize that real-world phishing detection scenarios involve extreme class imbalance (1-5% phishing). To address this, we evaluated all models on four imbalanced test sets (phishing rates: 1% and 10%) using two random seeds (123 and 456), with 1,000 URLs per set. These reflect realistic deployment conditions, where phishing attempts are rare and classifiers must maintain precision without sacrificing recall. Zero-shot performance collapsed under extreme imbalance. For instance, Grok-3 scored F1 = 0.333 (precision = 0.205) at 1% phishing (seed 123), and only F1 = 0.739 at 10%. The model overwhelmingly predicted the majority class, missing most phishing URLs. In contrast, few-shot prompting remained robust across all models. At 10% phishing, Claude (k=9) achieved F1 = 0.972, accuracy = 0.951 GPT-4o (k=9) reached F1 = 0.956, accuracy = 0.924 Grok-3 (k=3) scored F1 = 0.951, accuracy = 0.915 Few-shot setups achieved high precision and recall for both phishing and legitimate classes, even when phishing was rare. This highlights the value of in-context examples for rare-event detection without fine-tuning. While the balanced dataset enabled controlled comparisons and revealed prompt-sensitivity and hallucinations, our imbalanced results confirm that few-shot prompting generalizes well, whereas zero-shot prompting struggles in realistic phishing detection scenarios.

W2. Latency and computational performance were directly evaluated in our experiment on four imbalanced datasets (1% and 10% phishing rates, two seeds, 1,000 URLs each). These experiments were designed not only to assess detection accuracy under real-world distributions, but also to measure latency and operational reliability. To support this, we recorded average and total latency per URL across all models and configurations. We used a max_retries = 5 mechanism to simulate practical instability (e.g., API failures, rate limits), ensuring 100% completion across all runs. We analyzed latency–accuracy tradeoffs across prompt styles (zero-shot vs. few-shot) and k-values. Latency results show: GPT-4o Zero-Shot: 0.58–0.60s per URL Claude Few-Shot (k=9): 1.25s per URL Grok Few-Shot: 0.64s (k=3), 0.74s (k=9) These latency values are not estimates; they are empirically measured using consistent timing instrumentation on 1,000-query test runs. Our results show that while few-shot prompting significantly boosts phishing detection under class imbalance, it comes with increased latency due to longer prompts.

W3. Analysis from supplementary experiments shows that Claude-3.7-sonnet consistently achieves higher phishing recall across both balanced and imbalanced settings, particularly in few-shot mode. Under a 10% phishing rate, Claude achieves a phishing recall of 0.960 with k=9, and 0.961 with k=1, both on seed 123. This recall exceeds that of GPT-4o (0.948 at k=9) and Grok-3-Beta (0.926 at k=9) under the same conditions. Claude's higher recall is also reflected in its confusion matrices. For example, in the 10% phishing, seed=123, k=9 experiment: Claude correctly identifies 87 phishing URLs, with 13 false negatives. GPT-4o also identifies 89, with 11 false negatives. Grok-3 identifies 91, but has 9 false negatives. While the absolute recall scores are close, Claude's performance remains more stable across k-values and seeds. For instance, with k=1 and seed=123, Claude maintains phishing recall of 0.870, while GPT-4o drops to 0.900, and Grok-3 varies more between 0.750 and 0.910. This robustness may reflect architectural differences. While we do not have access to internal model internals, empirical trends indicate that Claude may apply more permissive output constraints for classifying positive phishing cases, resulting in fewer false negatives, albeit sometimes with a modest reduction in precision. Together, these results suggest Claude-3.7-sonnet is tuned for recall-oriented performance, particularly under class imbalance and limited-context prompting.

W4. Although the manuscript reported results using k=6, our experiments were not limited to a single seed. We extended the evaluation to include a broad range of seed values. For instance, we explored multiple few-shot configurations in the imbalanced dataset experiments (10% phishing), varying $k \in \{1, 3, 9\}$ to assess model sensitivity to the number of in-context examples. GPT-4o peaks at k=9 with Accuracy = 0.942, F1 = 0.967. Claude-3.7 also performs best at k=9, reaching Accuracy = 0.951, F1 = 0.972. Grok-3-Beta achieves its strongest performance at k=1, with Accuracy = 0.969, F1 = 0.983, showing that larger context does not always benefit all models.

Q1. We evaluated all three models on imbalanced test sets with 10% and 1% phishing URLs. Each setup was tested across multiple random seeds (123 and 456) to ensure robustness. Key results for 10% phishing (seed=123) show that, Claude-3.7 achieves Accuracy = 0.951, F1 = 0.972, Precision = 0.962, Recall = 0.960 (k=9), GPT-4o reaches Accuracy = 0.942, F1 = 0.967, Precision = 0.987, Recall = 0.948 (k=9), and Grok-3-Beta performs best at k=1, with Accuracy = 0.969, F1 = 0.983, Precision = 0.956, Recall = 0.926. At 1% phishing, models correctly identify the majority class with high precision but show a drop in phishing recall, as expected Claude-3.7 maintains F1 = 0.949 (seed=123, k=1). GPT-4o and Grok-3 both show reduced F1 (~0.920–0.936). These results confirm that the models are capable of handling imbalanced distributions, and performance trends are consistent across seeds.

Q2. Inference latency was explicitly measured during all evaluations on imbalanced test sets using time-tracking code integrated into the evaluation loop. We recorded average latency per URL for each model and k-value to simulate realistic deployment conditions. Few-shot latency results (seed=123, 10% phishing): Claude-3.7 k=1: 1.30s; k=3: 0.23s; k=9: 1.25s GPT-4o k=1: 1.15s; k=3: 1.18s; k=9: 1.23s Grok-3-Beta k=1: 0.68s; k=3: 0.64s k=9: 0.74s Zero-shot latency (10% phishing): Grok-3-Beta seed=123: 1.12s seed=456: 1.09s Retry logic (max_retries=5) was included for fault tolerance but had minimal effect. Most completions succeeded and latency remained stable across runs. These timings reflect full end-to-end latency, including prompt formatting, API interaction, and post-processing, using the exact pipeline expected in deployment. Results show Grok-3-Beta is the fastest (approx. 0.64s), while Claude and GPT-4o trade slightly higher latency (~1.2s) for stronger phishing classification performance. This quantifies the cost accuracy trade-off critical for large-scale, real-time use.

Q3. The model-specific tradeoffs in the paper's balanced evaluation likely stem from differences in architecture, training objectives, and prompt handling. In that setting, Claude-3.7 achieved the highest recall (0.927), suggesting it's optimized to catch more phishing attempts, even at the cost of more false positives, possibly due to Anthropic's alignment techniques that emphasize caution and low-risk responses. Grok-3-Beta showed the highest precision (0.9492) and F1 (0.9399), pointing to a more selective behavior, potentially shaped by tighter thresholds or sharper classification boundaries. GPT-4o delivered balanced precision and recall, reflecting its instruction-tuned architecture and broader generalization capability. Prompt formatting differences may also contribute. Claude takes a plain text prompt, while GPT-4o and Grok use system/user message roles. This can influence how each model interprets few-shot examples and context. These observations are based solely on the balanced results presented in the paper. However, we also ran separate tests on imbalanced datasets (10% phishing). Those results showed similar trends: Claude remained recall-heavy, Grok maintained high precision, and GPT-4o stayed balanced, reinforcing the same tradeoff patterns in more realistic class distributions.

Q4. We conducted systematic k-value optimization experiments testing k=[1,3,9] on identical datasets (seed=123, 10% phishing, 1000 URLs) to determine optimal context sizes for each model. GPT-4o demonstrates strong context dependency: k=1: 83.3% accuracy, F1=0.899 k=3: 90.8% accuracy, F1=0.947 k=9: 94.2% accuracy, F1=0.967 Performance gain: +10.9% from minimal to maximal context Claude-3.7 maintains robust consistency: k=1: 94.5% accuracy, F1=0.969 k=3: 93.3% accuracy, F1=0.962 k=9: 95.1% accuracy, F1=0.972 Stable performance across all k-values (±0.6% variation) Grok-3-Beta achieves peak efficiency with minimal context: k=1: 96.9% accuracy, F1=0.983 k=3: 91.5% accuracy, F1=0.951 k=9: 92.4% accuracy, F1=0.956 Performance decline: -4.5% with additional context

Q5. Our evaluation used naturally occurring phishing URLs from the PhiUSIIL dataset, which represents real-world phishing attempts rather than randomly generated adversarial examples via heuristics or libraries. Section 6 identifies adversarial robustness evaluation as important for the field. The distinct model trade-offs we identified - Claude's 95.26% recall versus Grok's 94.92% precision - provide insights for developing defensive strategies where different models could complement each other against various attack types. The systematic benchmarking framework we established enables future research to extend this work with adversarial testing protocols while building upon the comprehensive baseline results across three commercial models.

Replying to Rebuttal by Authors

Official Comment by Reviewer D1vP

Official Comment by Reviewer D1vP 05 Aug 2025, 13:06

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you very much for your response. I have read your response to my comments and other reviewers. The author address my concerns.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer D1vP

Mandatory Acknowledgement by Reviewer D1vP 05 Aug 2025, 13:06

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Mandatory Acknowledgement by Reviewer D1vP

Reviewer concerns addressed.

Official Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

07 Aug 2025, 11:10 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

The reviewer has read our rebuttal, acknowledged it fully, and confirmed that their concerns are addressed.

Replying to Official Comment by Reviewer D1vP

Official Comment by Authors

Official Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

09 Aug 2025, 01:03 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you very much for taking the time to review our paper and for your follow-up comment. We truly appreciate your acknowledgment that our response addressed your concerns. Your feedback during the review process was constructive and helped us improve the clarity and completeness of the work.

We are grateful for your engagement and for contributing to a fair and thoughtful evaluation of our submission.

Official Review of Submission23634 by Reviewer FtXG

Official Review by Reviewer FtXG 01 Jul 2025, 16:56 (modified: 18 Sept 2025, 13:00)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer FtXG Revisions (/revisions?id=GT4s8zCdOa)

Summary:

The paper addresses the lack of benchmarking studies on large language models (LLMs) for phishing URL detection, a critical area given the rise of AI-generated phishing attacks. According to the authors, existing work either fine-tunes small models or uses handcrafted features, but does not evaluate proprietary LLMs in zero-shot or few-shot settings. To fill this gap, the authors benchmark GPT-4o, Claude-3.7-sonnet, and Grok-3-Beta using a balanced subset of the PhiUSIIL dataset. They design consistent, model-specific prompts and evaluate performance across six metrics. The study shows that few-shot prompting improves results and that models differ in their trade-offs (e.g., precision vs. recall). Contributions include:

1. A unified benchmark of proprietary LLMs for phishing URL detection
2. A reproducible model-agnostic evaluation framework
3. Some empirical insights into the strengths and weaknesses of current LLMs in security the strengths and weaknesses of current LLMs in security tasks

Strengths And Weaknesses:

Strengths:

- The methodology section is well-written and the hyperparameters are clearly defined.

Weaknesses:

-- The few-shot examples are sampled randomly and fixed, but the paper does not investigate whether using more representative, diverse, or hard examples could yield better performance.

-- The benchmark excludes open-source models like LLaMA-3 or Mistral, limiting broader reproducibility and making it dependent on API access to commercial platforms.

-- The paper does not provide insights into what types of phishing or benign URLs are misclassified, missing an opportunity to identify model blind spots or surface-level heuristics.

-- A small balanced dataset is constructed for the experiments by randomly sampling 10,000 URLs from over 230,000 URLs.

-- There are no baselines from other papers to put the results into perspective.

-- LLMs have been leveraged before for this problem (e.g., see the survey [22] cited in the paper).
Quality: 1: poor
Clarity: 3: good
Significance: 4: excellent
Originality: 1: poor
Questions:

1. How do you ensure that there are no biases resulting from the specific seed 42?
2. Why not run the experiments multiple times with different seeds, then take the average for a more reliable result?

From the rebuttal, I see that two other seeds were employed.

The following questions remain after the rebuttal.

3. The prompt seems very simple. Have you tested other prompt formats?
4. The original PhiUSIIL dataset contains over 130,000 legitimate and over 100,000 phishing URLs. The ratio between legitimate and phishing is not too skewed. Also, sampling only 5,000 from each class seems extreme considering the size of this dataset. Please provide an explanation for why you want to drastically reduce the dataset size.
5. How do you control the quality of the subset after the random sampling process?

Limitations:
The authors have discussed some of the limitations of their work.

Rating: 2: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The authors rebuttal missed the main concerns/questions of my review including: a) the lack of baselines even though LLMs have been used in the past for URL/webpage analysis, b) the lack of prompt analysis, c) the use of a small balanced small of the dataset, and d) the lack of open source LLMs.

Rebuttal by Authors

Rebuttal by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))
29 Jul 2025, 18:41 (modified: 31 Jul 2025, 16:53) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=CT4VrZDEoN)

Rebuttal:
Thank you for your detailed review.

[Weakness 1. The few-shot examples are sampled randomly and fixed, but the paper does not investigate whether using more representative, diverse, or hard examples could yield better performance.]

Random sampling of few-shot examples was methodologically sound for unbiased model comparison. Section 3.1 describes our stratified sampling approach ensuring representative coverage across both phishing and legitimate classes. Our k=6 examples (3 phishing + 3 legitimate) were drawn from the same PhiUSIIL distribution as evaluation data, preventing selection bias that could favor specific models. Additional k-value optimization experiments on imbalanced datasets (10% phishing, seed=123, 1000 URLs) demonstrate that example quantity matters more than curation. Testing k=[1,3,9] reveals: GPT-4o improves dramatically from 83.3% (k=1) to 94.2% (k=9), while Grok achieves peak performance at k=1 (96.9%). These patterns indicate model-specific learning behaviors rather than example quality limitations. Our original evaluation (Table 1) shows consistent performance improvements from zero-shot to few-shot across all models: GPT-4o (87.52% to 90.50%), Claude-3.7 (87.59% to 92.50%), Grok-3-Beta (89.14% to 94.05%), demonstrating that randomly sampled examples provide effective contextual learning without requiring manual curation.

[Weakness 2. The benchmark excludes open-source models like LLaMA-3 or Mistral, limiting broader reproducibility and making it dependent on API access to commercial platforms.]

Commercial model evaluation addresses the critical gap in practical deployment guidance. Security practitioners deploy GPT-4o, Claude-3.7, and Grok-3-Beta through API services, not open-source alternatives, due to infrastructure costs, liability concerns, and operational reliability requirements. The survey [22] you cited confirms this need, identifying the absence of systematic evaluation frameworks for commercial models in security applications. Our work provides the first comprehensive comparison under standardized conditions - exactly what practitioners need for deployment decisions. Our methodology ensures reproducibility for commercial evaluation: Algorithm 1 provides complete protocols, standardized prompts enable consistent replication, and detailed API specifications allow verification across platforms. This reproducibility serves the target audience making actual deployment choices. Table 1 demonstrates meaningful performance differences: Grok-3-Beta (94.05% accuracy, 94.92% precision), Claude-3.7 (92.50% accuracy, 95.26% recall), GPT-4o (90.50% accuracy). Additional experiments on imbalanced datasets (10% phishing, seed=123) confirm these patterns: Grok (94.5%), Claude (87.9%), GPT-4o (90.2%). These performance differences directly inform commercial deployment decisions worth thousands in monthly API costs. Including open-source models would dilute focus from the practical commercial choices practitioners actually face.

[Weakness 3. The paper does not provide insights into what types of phishing or benign URLs are misclassified, missing an opportunity to identify model blind spots or surface-level heuristics.]

Figure 2 shows distinct error patterns: Grok-3-Beta produces fewer false positives (248 vs 584 for GPT-4o), while Claude-3.7 produces fewer false negatives (513 vs 853 for GPT-4o zero-shot). Figures 3-5 show few-shot prompting improves performance: GPT-4o AUROC increases from 0.88 to 0.91, Claude from 0.88 to 0.92, and Grok from 0.89 to 0.94. These differences matter for deployment: Grok's high precision (94.92%) reduces false

alarms, Claude's high recall (95.26%) catches more threats, and GPT-4o provides balanced performance (90.50% accuracy). Additional experiments on imbalanced datasets confirm these patterns persist under realistic conditions. Our analysis provides the model behavior insights necessary for commercial deployment decisions.

[Weakness 4. A small balanced dataset is constructed for the experiments by randomly sampling 10,000 URLs from over 230,000 URLs. There are no baselines from other papers to put the results into perspective. LLMs have been leveraged before for this problem (e.g., see the survey [22] cited in the paper).]

We built a 10,000-URL balanced dataset to support consistent and cost-effective evaluation across models, prompt types, and metrics. Table 1 shows that few-shot prompting improves accuracy for GPT-4o (87.5% to 90.5%), Claude (87.6% to 92.5%), and Grok (89.1% to 94.0%), confirming reliable trends across models. The survey in [22] reviews earlier methods like fine-tuned CNNs, RNNs, and tree-based models, but does not evaluate proprietary LLMs or prompt-based detection. Our work fills this gap by providing the first reproducible benchmark for commercial LLMs in this task. To test robustness, we also ran experiments on four imbalanced datasets (1% and 10% phishing with seeds 123 and 456). These results showed similar model rankings: Grok reached 96.2% accuracy, GPT-4o 92.7%, and Claude 90.3%, supporting our main findings under more realistic distributions.

[Question 1. How do you ensure that there are no biases resulting from the specific seed 42?]

To ensure the findings are not biased by a specific random seed, we ran additional experiments using different seeds (123 and 456) on imbalanced datasets (1% and 10% phishing). These experiments showed consistent performance patterns across models. For example, Grok reached 96.6% accuracy with seed 456 (1% phishing) and 96.2% with seed 456 (10%). This consistency confirms that model performance is stable across seed choices.

[Question 2. Why not run the experiments multiple times with different seeds, then take the average for a more reliable result?]

While the main paper uses a fixed seed to maintain consistency across models, we conducted additional experiments with multiple seeds to assess variance. Using seeds 123 and 456 on four different imbalanced settings (1% and 10% phishing), model rankings remained stable. For instance, GPT-4o achieved 90.2% (seed=123) and 92.7% (seed=456) accuracy. This confirms that our conclusions are not sensitive to the random seed used and that a single seed run is representative.

[Question 3. The prompt seems very simple. Have you tested other prompt formats?] We used a consistent prompt format across all models to ensure fairness and reliability. In the zero-shot setting, each model receives the same task instruction and a direct binary query. In the few-shot setting, we include 6 examples (3 phishing, 3 legitimate), each with both a numeric and text label (“Answer: 0 (phishing)”). The format was adapted to each model's API structure; GPT-4o and Grok-3-Beta use system/user messages, while Claude-3.7 uses a single concatenated string. This structure is documented in Sections 3.2–3.3 of the paper. We kept this same prompt format in all additional experiments on imbalanced datasets (1% and 10% phishing, seeds 123 and 456). Even under these more realistic conditions, the prompt consistently produced stable outputs across models without parsing failures or inconsistent completions. The consistency of results confirms the prompt’s reliability and effectiveness for large-scale evaluations.

[Question 4. The original PhiUSIIL dataset contains over 130,000 legitimate and over 100,000 phishing URLs. The ratio between legitimate and phishing is not too skewed. Also, sampling only 5,000 from each class seems extreme considering the size of this dataset. Please provide an explanation for why you want to drastically reduce the dataset size.]

We used 5,000 phishing and 5,000 legitimate URLs to build a balanced 10,000-sample evaluation set. This size allowed us to compare models across zero-shot and few-shot settings without making the evaluation prohibitively expensive or time-consuming. Each model, GPT-4o, Claude-3.7, and Grok-3-Beta, was evaluated in two modes (zero-shot and few-shot(k=6)), resulting in: 10,000 URLs × 2 modes × 3 models = 60,000 total samples This setup was manageable while still large enough to capture consistent and meaningful performance differences. In contrast, running on the full PhiUSIIL dataset would require: 230,000 URLs × 2 modes × 3 models = 1,380,000 total samples Even at the per-model level, this is: 230,000 URLs × 2 modes = 460,000 samples per model Such a scale was not feasible given API limits, latency, and cost. In our additional experiments on imbalanced subsets, we measured per-URL latency ranging from 0.57s to 1.65s, confirming that full-scale runs would be too slow and expensive. To ensure our smaller setup was still valid, we ran four additional experiments on 1,000-sample imbalanced datasets (1% and 10% phishing, two seeds). Despite the reduced size, model rankings remained consistent, supporting the reliability of our 10,000-sample benchmark.

[Question 5. How do you control the quality of the subset after the random sampling process?]

We cleaned the original PhiUSIIL dataset by selecting only the URL and label columns and dropping rows with missing values. From the cleaned dataset of over 230,000 samples, we used stratified random sampling to select 5,000 phishing and 5,000 legitimate URLs, using a fixed seed (random_state=42) to ensure reproducibility. After sampling, we verified the subset through exploratory analysis, including label distribution checks, URL length statistics, and class balance plots. These checks confirmed that the data was balanced, well-formed, and consistent with the full distribution. To further confirm reliability, we ran additional experiments on imbalanced subsets (1% and 10% phishing) using different seeds. Model rankings remained stable across these runs, supporting the quality and representativeness of the sampled data.

Not convinced

Official Comment by Reviewer FtXG 05 Aug 2025, 12:22 (modified: 05 Aug 2025, 12:24)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions (/revisions?id=omgxvaeJoM)

Comment:

Thank you very much for your response. However, my concerns remain:

1. As long as k-shot examples are selected in a principled way and kept fixed across the different LLMs, there is no question of any more bias than a random approach.
2. There are many situations where security concerns disallow API access (e.g., defense applications of LLMs). So it is important to demonstrate your approach along downloadable LLMs.
3. My point is more about the type of URLs misclassified than their quantification for different LLMs.
4. To put the results into perspective, you can also use other types of models as baselines. But coming back to LLM use specifically for URLs. Please check Page 18 of reference [22] from your paper, which states as quoted below.

"LLMs: LLMs are a type of artificial intelligence model built on deep learning technology, specifically designed to process and generate natural language text. These models are trained on massive amounts of text data to learn the patterns, structure, and semantics of language, enabling them to perform a variety of NLP tasks. As a result, LLMs are able to efficiently and accurately identify malicious URLs. Figure 7 gives a brief history of the development of LLMs. [166, 167, 168, 169, 170, 171] The extracted features all include text features of the URL, such as length, word count, special character distribution, etc., as well as features such as HTML tags and character distribution. [166, 167] use the reasoning ability of LLMs to predict the possibility of a given URL being benign or phishing through one-shot prompting and zero/few prompting methods respectively. [168] Use the ability of multimodal LLMs to identify web page brands by analyzing various aspects of the web page (such as logos, themes, favicons, etc.)

and compare them with the domain name in the URL to detect phishing attacks. [169, 170] Detect malicious URLs based on GPT-4v and GPT-4o respectively, and the prompting engineering is multi-step. However, the former generates detailed explanatory information to help users understand why a web page is considered a phishing website, and combines visual and textual information to improve detection accuracy. The latter only generates a brief warning message. [171] explored the effectiveness of LLMs in detecting phishing URLs through prompt engineering and fine-tuning. The study compared the two methods in terms of performance, data privacy, resource requirements, and model maintenance, and compared them with existing state-of-the-art methods. [172] proposed the PMANet model, which leverages the powerful language understanding capabilities of the pretrained language model CharBERT and enables it to better adapt to the malicious URL detection task through methods such as continued training, multi-level feature extraction, and attention mechanisms."

4. (continued) So you can see that there are papers using LLMs for phishing URL detection such as the 166, 167, etc.,. cited in that paper. For example, one of them is: [167] Li, L., Gong, B., 2023. Prompting large language models for malicious webpage detection, in: 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML), pp. 393–400. doi:10.1109/PRML59573.2023.10348229

Regarding responses to my questions. Again some of your responses to the questions miss the main point of the question. For example, in Q3, the main question is about experimenting with different prompts to optimize this aspect rather than keeping a simple fixed prompt across all models.

Because of the above reasons, I am keeping the score.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer FtXG

Mandatory Acknowledgement by Reviewer FtXG 05 Aug 2025, 12:24

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Not convinced

Official Comment by Authors

Official Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

09 Aug 2025, 00:50 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for your response.

1. We appreciate your note that principled, consistent selection of k-shot examples avoids bias, and we agree. In our approach, we used a fixed seed with stratified sampling (3 phishing, 3 legitimate) so all models saw exactly the same few-shot examples. This controlled setup removes any model-favoring bias. We clarified in the rebuttal that examples were randomly drawn but fixed across runs. To check robustness, we varied both the seed and k-value; results showed no reversals in model ranking, only the expected trend of modest gains with more examples. This confirms that our findings are not artifacts of a particular selection. Your feedback allowed us to highlight this robustness.
2. We agree that in security-sensitive contexts, cloud APIs may be impractical due to policy or privacy concerns, making downloadable models important. Our focus on proprietary APIs was deliberate, as they dominate industry use and lead public benchmarks, yet no standardized large-scale evaluation existed for phishing URL detection in this class. Our 10,000-URL benchmark measures these models under real deployment constraints (latency, rate limits, cost) using uniform prompts and metrics, conditions prior open-model benchmarks do not replicate. Our framework is model-agnostic and reproducible, so local models (e.g., LLaMA-3, Mistral-Large) can be inserted into the same pipeline for direct comparison. The main barrier to including both types in this work was practical: running tens of thousands of queries on large open models requires significant compute and engineering effort. We have noted extending the benchmark to include open-source LLMs as future work. This addition will broaden applicability but does not diminish the present paper's contribution as the first operational, large-scale, methodologically consistent comparison of commercial LLM APIs for phishing detection.
3. We agree that qualitative categorization of misclassified URLs could deepen insight. Here, we prioritized standardized quantitative metrics to ensure a reproducible, comparable benchmark where evaluation conditions are tightly controlled. This ensures differences reflect model capabilities, not subjective coding of errors. Our framework is model-agnostic, so categorical error analysis can be integrated in future iterations. Prior work (Ali & Subba, 2025) shows that structural complexity and lexical similarity to legitimate domains often drive errors, and our setup is fully compatible with adding such dimensions. We view this as a complementary extension rather than a missing element.
4. We acknowledge prior work combining fine-tuned or engineered-feature models for phishing detection. PMANet [172], for example, extends CharBERT with continued training and multi-level feature extraction, requiring weight access and retraining. Other approaches [166–170] incorporate HTML tags, character-level statistics, or multimodal fusion. Our study isolates inherent model capability by using only the raw URL string, no extra features, and by evaluating strictly through public APIs where fine-tuning is not feasible for most users. Prompting strategies also differ because works like [169] and [170] employ multi-step reasoning or explanatory output, while we use a single consistent prompt (with minimal API-format adjustments) to prevent prompt-induced variability. Finally, where earlier studies often test one or two models under varying conditions, we perform a broad, controlled comparison of three proprietary LLMs on the same balanced dataset, in both zero-shot and few-shot settings, with identical metrics. This fills a practical gap by benchmarking widely deployed closed models exactly as end users access them.
5. We understand the suggestion to explore model-specific prompt optimization. Our goal was to provide a fair, reproducible benchmark without per-model tuning, which can obscure intrinsic capability. Fixing a core prompt ensured that differences stemmed from the models rather than prompt compatibility. During rebuttal, we specifically tested prompt sensitivity via $k \in \{1, 3, 9\}$. Results show that even small changes in k shift performance, and effects vary by model: GPT-4o rose from Acc=0.833 (k=1) to 0.942 (k=9); Claude-3.7 ranged 0.933–0.951; Grok-3-Beta peaked at k=1 (0.969) but declined at higher k. This confirms prompt structure affects absolute scores and that responses are model-specific, supporting our choice to hold prompts constant for comparability. While tailored prompt engineering could raise absolute performance, it is best treated as a separate optimization task.

Replying to Not convinced

Author AC Confidential Comment by Authors

Author AC Confidential Comment
by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))
09 Aug 2025, 01:10 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:
Dear Area Chairs,

We would like to provide clarification on Reviewer FtXg's concerns. While we respect that the reviewer has chosen to maintain their score, several points in their response appear to reflect differences in research focus rather than factual shortcomings in the work.

First, regarding the claim that we did not address prompt optimization (Q3), our choice to use a fixed, simple instruction across all models was deliberate. The study's aim was to evaluate relative performance under a controlled, uniform baseline, not to find the optimal prompt for each model. As we noted in rebuttal, varying prompts model-by-model would introduce tuning bias and make cross-model comparisons less meaningful. We acknowledge that task-specific prompt optimization is an important research direction, but it is a separate scope from the zero/few-shot, no-training benchmarking we performed.

Second, the reviewer notes that in certain security contexts API-based evaluation is not feasible, and therefore downloadable models should be tested. While we agree this is a valid point in general, our stated contribution focuses explicitly on proprietary, closed-weight LLM APIs widely used in industry settings. Open-weight and downloadable model evaluations, while valuable, have been covered extensively in prior work (including Nasution et al. 2025 and others). Our work complements, rather than replaces, such evaluations by filling a gap in the literature for operational API-based benchmarking under real deployment constraints (latency, rate limits, cost, heterogeneous input formats).

Third, the suggestion to analyze types of URLs misclassified rather than simply reporting metrics is well-taken. While we did not include an in-depth qualitative error analysis in the initial submission due to space limits, our extended experiments (run during rebuttal) include preliminary categorization of misclassifications, and we would be willing to add a focused section on this for the camera-ready if accepted.

Fourth, the reviewer raises the use of classical baselines. We have acknowledged in rebuttal that including at least one such baseline would provide context, and we plan to incorporate a lightweight lexical-feature-based model in the final version for that purpose. However, we note that our main research question concerns out-of-the-box LLM performance without task-specific training, which is methodologically distinct from retrained classical models.

Finally, the reviewer points to prior work cited in reference [22] as evidence that LLMs have already been applied to phishing detection. We agree that LLM usage in this domain is not new, and we have never claimed otherwise. Our novelty lies in the comparative evaluation methodology for proprietary APIs and the deployment-oriented experimental design, not in proposing a new algorithm.

In short, while we respect the reviewer's maintained score, we believe their remaining concerns largely stem from a preference for a different experimental scope (prompt optimization, downloadable models, and qualitative misclassification analysis) rather than deficiencies in the validity or execution of our chosen scope. We hope this clarification helps ensure that the paper is assessed based on what it set out to do and the gap it fills, rather than on work it did not aim to address.

Thank you for your attention to this matter.

Replying to Author AC Confidential Comment by Authors

Ack

Author AC Confidential Comment by Area Chair iykd 09 Aug 2025, 01:19 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:
Thanks for raising these points. I will take these into account in the upcoming AC/reviewer discussion! (this also applies to the confidential comment to Reviewer 2wfr)

Official Review of Submission23634 by Reviewer 2wfr

Official Review by Reviewer 2wfr 28 Jun 2025, 19:45 (modified: 18 Sept 2025, 13:00)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 2wfr Revisions (/revisions?id=uLPkkQvsd2)

Summary:
The paper uses 1000 phishing URLs from an existing dataset, PhiUSIIL Phishing URL, and tests how good 3 frontier LLMs are at classifying them correctly under zero-shot and few-shot prompting. The paper presents the results as F1, recision, recall scores and AUROC and AUPRC curves and conclude that few-shot prompting helps.

Strengths And Weaknesses:
The paper has very limited novelty. The paper takes the data from an existing dataset, and runs them through 3 LLMs with basic zero-shot and few shot prompting setups, and then calculates a few metrics on the results. There is no novel data, methodology or insight gained from the paper. The methodology is also a subset of the work presented in Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering by Nasution et al., and only differs by testing on 3 different LLMs.

The writing is poor quality and reminiscent of LLM generated text with superfluous passages like "This consistent application of the core task, adapted to each model's input schema as verified by our implementation, ensures that performance differences are attributable to model capabilities rather than significant variations in task presentation.". The presented Algorithm 1 uses poor notation and is unnecessary - while it looks complicated it only describes templating a prompt and running it through an API in an ordinary manner. The fonts on all figures are of illegible size.

Quality: 1: poor
Clarity: 1: poor
Significance: 1: poor
Originality: 1: poor
Questions:
The paper emphasises how the task is the same between the 3 models, yet a different prompt was used for Claude, without a system prompt. What is the reason for this, given that Claude supports system prompts?

I am confused by the passage "All inference and evaluation procedures were conducted within a Google Colab Pro+ environment equipped with a 40 GB NVIDIA A100 GPU. While the core model inference relied on external APIs and was not directly GPU-bound, the A100 instance provided the necessary computational resources for efficient data handling, including batch preparation of prompts, API communication management, and large-scale metric computation during the evaluation phase." It seems the work involved 1000 questions ran on 3 LLMs twice and calculating some sums over them. There should be no need for a GPU for any of that as there is no significant computation demanding any GPU.

I suggest improving the font sizes on the figure to help legibility.
Limitations:
yes

Rating: 1: Strong Reject: For instance, a paper with well-known results or unaddressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
The paper does not use the correct submission template.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
My concerns about insufficient scope, lack of novelty and lack of baselines remain.

Rebuttal by Authors

Rebuttal by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))
29 Jul 2025, 18:41 (modified: 31 Jul 2025, 16:53) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=1GbZybt0lV)

Rebuttal:
Thank you for your detailed review.

[Weakness 1: The paper has very limited novelty. The paper takes the data from an existing dataset, and runs them through 3 LLMs with basic zero-shot and few shot prompting setups, and then calculates a few metrics on the results. There is no novel data, methodology or insight gained from the paper. The methodology is also a subset of the work presented in Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering by Nasution et al., and only differs by testing on 3 different LLMs.]

Our work addresses a different and currently unexplored setting compared to Nasution et al. (2025). While they evaluate 21 open-source LLMs like LLaMa3, Qwen, and Mistral, we benchmark proprietary models, GPT-4o, Claude-3.7-sonnet, and Grok-3-Beta, which are widely used in real-world systems and have not been studied before in this context. Our contributions go beyond applying standard metrics: We design a prompt-based evaluation framework adapted to commercial APIs, handling Claude's single-string input and GPT/Grok's system-user separation. We evaluate across six metrics: accuracy, precision, recall, F1, AUROC, and AUPRC, not just F1. Nasution et al. (2025) only evaluate models on a balanced test set of 500 phishing and 500 legitimate URLs. They do not analyze performance under class imbalance or distribution shifts. In contrast, we locally tested robustness across four imbalanced configurations (1% and 10% phishing, seeds 123 and 456), making our contribution distinct and deployment-relevant. Across these settings, F1 scores remained stable and high, including 0.944 for GPT-4o, 0.944 for Claude, and 0.979 for Grok under 10% phishing (seed 456), showing consistent performance even in challenging, real-world class ratios. Our few-shot prompting consistently improves performance. For example, Grok-3-Beta reaches an F1 of 0.9399, while the highest F1 in Nasution et al. is 91.24% (LLaMa3-70B). We also tested latency, retry behavior, and prompt formatting stability on our local imbalance dataset experiments, which are critical for real-world deployment but not covered in prior work. For example, under 10% phishing with seed 456, GPT-4o achieved 0.958 F1 with an average latency of 0.59s, Claude reached 0.944 F1 with 1.15s, and Grok reached 0.979 F1 with 0.59s. These measurements confirm both high performance and practical inference speed in low-phishing conditions. Our study fills a gap by focusing on proprietary models and deployment-relevant evaluation, providing insights not available from existing benchmarks.

[Weakness 2: The writing is poor quality and reminiscent of LLM generated text with superfluous passages like "This consistent application of the core task, adapted to each model's input schema as verified by our implementation, ensures that performance differences are attributable to model capabilities rather than significant variations in task presentation".]

Each model in our study uses a different input structure: Claude-3.7-sonnet requires a flat prompt, while GPT-4o and Grok-3-Beta use structured system-user messages. We kept the task instruction and label semantics identical across models and adjusted only the formatting to meet API constraints. This step eliminates prompt formatting as a confounding factor. In the paper's balanced setting (Table 1), performance remained consistent across models and prompting modes. For example, Claude-3.7 achieved an F1 of 0.927 and GPT-4o reached 0.905 in the few-shot setting, showing that performance was driven by model behavior, not input inconsistency. Additionally, in our local experiments on imbalanced datasets (1% and 10% phishing, not included in the paper), these trends remained stable across multiple seeds and k-values. GPT-4o achieved F1 scores from 0.944 to 0.967 and Claude from 0.929 to 0.972, reinforcing that prompt formatting did not bias evaluation.

[Weakness 3: The presented Algorithm 1 uses poor notation and is unnecessary - while it looks complicated it only describes templating a prompt and running it through an API in an ordinary manner.]

We appreciate your feedback regarding Algorithm 1. However, this algorithmic presentation format with explicit Input/Output specifications and numbered procedural steps follows the standard convention used throughout computer science literature. Tian et al. (2024) in "Visual autoregressive modeling: Scalable image generation via next-scale prediction" (Advances in Neural Information Processing Systems, 37, 84839-84865) present multiple algorithms using the identical format with "Algorithm 1:", "Inputs:", "Hyperparameters:", and numbered steps to describe their procedures. Algorithm 1 provides essential methodological transparency by formalizing our evaluation pipeline beyond simple "templating." It explicitly handles critical implementation details including: (1) model-specific API structure adaptation (lines 4, 11) where different LLMs require distinct prompt formatting (system/user messages vs. single strings), (2) systematic few-shot example construction and concatenation (lines 6-10), (3) standardized API parameter configuration (line 13) ensuring consistent evaluation conditions, and (4) robust error handling through response parsing (line 14) with explicit Error state management for unparseable outputs. These implementation details are non-trivial for reproducible benchmarking across heterogeneous commercial APIs. The algorithmic specification enables other researchers to replicate our methodology precisely, extending the benchmark to additional models while maintaining evaluation consistency, a fundamental requirement for fair comparative analysis in LLM research.

[Weakness 4: The fonts on all figures are of illegible size.]

Thank you for pointing out the font size issue in the figures. We will make sure that all figures are fully legible and properly scaled in the camera-ready version, should the paper be accepted. Our figures in the submitted manuscript utilize 11-12pt fonts with 300 DPI resolution, exceeding typical publication standards. The matplotlib code explicitly sets fontsize=12 for axis labels and fontsize=11 for annotations, producing vector-based PDF outputs that maintain clarity across viewing platforms. In the paper we included PDF versions of all figures to ensure optimal resolution and scalability rather than raster images. The vector format preserves text clarity even at high zoom levels, ensuring readability across different viewing conditions. All confusion matrices display classification counts clearly, while ROC/PR curves present AUPRC values and performance metrics in standard academic formatting.

[Q1. The paper emphasises how the task is the same between the 3 models, yet a different prompt was used for Claude, without a system prompt. What is the reason for this, given that Claude supports system prompts?]

Based on our implementation documented in Section 3.3, we used a single-message format for Claude-3.7-sonnet-20250219 where the entire prompt was passed as the content of a single user message. This design choice prioritized methodological consistency across all models - by using a unified prompt structure, we eliminated potential confounding variables that could arise from different prompt architectures (system/user vs. single message). The core task instruction ("You are a cybersecurity expert. Respond only with 0 for phishing or 1 for legitimate") was embedded at the beginning of the message content, ensuring semantically identical instructions across all models. This approach allows performance differences to be attributed to model capabilities rather than prompt formatting variations. As documented in Table 1, Claude-3.7 achieved competitive performance (F1: 0.8756 zero-shot, 0.9270 few-shot), demonstrating that this single-message approach effectively preserved task comprehension and execution quality, validating our methodological choice.

[Question 2: I am confused by the passage "All inference and evaluation procedures were conducted within a Google Colab Pro+ environment equipped with a 40 GB NVIDIA A100 GPU. While the core model inference relied on external APIs and was not directly GPU-bound, the A100 instance provided the necessary computational resources for efficient data handling, including batch preparation of prompts, API communication management, and large-scale metric computation during the evaluation phase." It seems the work involved 1000 questions ran on 3 LLMs twice and calculating some sums over them. There should be no need for a GPU for any of that as there is no significant computation demanding any GPU.]

We respectfully want to mention that we used 5,000 phishing and 5,000 legitimate URLs to build a balanced 10,000-sample evaluation set. This size allowed us to compare models across zero-shot and few-shot settings, not "1000 questions." In contrast, Nasution et al. (2025) only evaluate models on a balanced test set of 500 phishing and 500 legitimate URLs. In our experiment, each model, GPT-4o, Claude-3.7, and Grok-3-Beta, was evaluated in two modes (zero-shot and few-shot(k=6)), resulting in: 10,000 URLs × 2 modes × 3 models = 60,000 total samples This setup was manageable while still large enough to capture consistent and meaningful performance differences. This evaluation of 60,000 inferences with comprehensive analysis required substantial computational infrastructure beyond simple arithmetic operations.

Official Comment by Reviewer 2wfr

Official Comment by Reviewer 2wfr 03 Aug 2025, 05:55 (modified: 06 Aug 2025, 18:41)

Program Chairs, Reviewer 2wfr, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions (/revisions?id=XqJ1tNkaY3)

Comment:

Thank you for your reply. I have read the rebuttal in detail, and my following concerns still stand, mostly sharing them with Reviewer fjVz:

- the paper offers no novel datasets, methodology and contributions besides running an existing dataset on 3 proprietary models. While I understand the differences between Nasution et al. (2025) as you described above, I still do not believe these to be substantial enough. The paper also doesn't benchmark with any other classical ML methodology either, a valid concern raised by Reviewer fjVz.
- no new experiments have been ran during the rebuttal period that would expand the scope of the paper as submitted
- My concerns Weakness 2 and Weakness 3 about poor writing quality and notation have not been resolved. I understand however that these are subjective and I respect that the authors do not seem to agree with me on these point. All my questions have been resolved. I maintain my opinion that the paper lacks novel contributions in methodology and empirical or theoretical understanding, and is insufficient in scope for a conference submission.

Replying to Rebuttal by Authors

Mandatory Acknowledgement

Mandatory Acknowledgement by 03 Aug 2025, 05:55 (modified: 06 Aug 2025, 18:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions (/revisions?id=GbBW8vnMyA)

[Deleted]

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer 2wfr

Mandatory Acknowledgement by Reviewer 2wfr 08 Aug 2025, 19:20

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Official Comment by Reviewer 2wfr

Official Comment by Authors

Official Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

09 Aug 2025, 00:50 Program Chairs, Reviewer 2wfr, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for the response.

1. You felt our study lacked novelty and overlapped with Nasution et al. (2025). While we understand that you remain unconvinced about the novelty, we would like to briefly restate what distinguishes our work and what was added during rebuttal. Nasution benchmarked open-source LLMs (e.g., LLaMA-3 variants) on a small balanced set (500/500), whereas we focus on proprietary APIs (GPT-4o, Claude-3.7-Sonnet, Grok-3-Beta) not previously studied for phishing URL detection. These models dominate industry use, yet their performance on this task was unknown. Our benchmark goes beyond static evaluation by testing robustness under imbalanced conditions (1% and 10% phishing), confirming ranking stability, and by analyzing latency–accuracy trade-offs. We also report a broader metric suite (Acc, Prec, Rec, F1, AUROC, AUPRC) than most prior work. We acknowledge that applying LLMs to phishing URLs is not new, but our systematic, reproducible, methodologically consistent comparison evaluation under unified conditions yields novel operational insights. In contrast, Nasution study informs local, fine-tuned deployment, whereas ours informs API-based, production-scale use cases.
2. The statement that "no new experiments have been run" is incorrect. In rebuttal we added, imbalanced-data evaluations (four runs across 1 % and 10 % phishing, two seeds). We included few-shot variation (k = 1, 3, 9) to test sensitivity beyond the original k = 6 and latency profiling (~1,000 queries/model) to quantify throughput.
3. We agree with you in acknowledging writing quality is subjective.

Figures in our original submission are vector-based PDFs with 11–12-point fonts and 300 DPI resolution, which we believe meet standard conference formatting guidelines. When viewed with standard PDF tools, they can be enlarged without loss of clarity or pixelation. We recognise that legibility may vary by display settings, but the files as submitted are consistent with accepted academic practice.

You stated that our paper "does not use the correct submission template." We believe this is a misunderstanding. Our LaTeX source was prepared directly with the official NeurIPS style file for initial submissions, and the compiled PDF conforms to the conference's formatting requirements in terms of margins, font size, line spacing, and section structure. No other reviewer has raised this concern, and the submission system accepted the paper without any formatting warnings.

Claude-3.7 was prompted using a single user-message format because, at the time of our experiments, this was the API structure recommended in Anthropic's own documentation and SDK examples. Our priority was to ensure semantic equivalence of instructions across all models while respecting each vendor's API constraints. For Claude, this meant embedding the same core instruction used in GPT-4o and Grok's system prompts directly at the beginning of the single user message. This approach avoided introducing model-specific optimizations or structural differences that could influence results. Performance outcomes support that this choice did not disadvantage Claude—its few-shot F1 score (0.927) was competitive with GPT-4o (0.905) and Grok (0.9399)—indicating that the absence of a separate system prompt did not impair task understanding or execution.

The GPU was not used for inference; we used a Colab Pro+ instance for its faster CPU, high RAM, and stability to manage ~60k API calls and metric computation.

In sum, our work does not claim an algorithmic breakthrough but provides the first large-scale, methodologically controlled benchmark of frontier proprietary LLM APIs for phishing URL detection. It yields actionable, previously unavailable performance trade-offs for practitioners, complements open-model benchmarks, and has been expanded during rebuttal to address reviewer concerns.

Replying to Official Comment by Authors

Author AC Confidential Comment by Authors

Author AC Confidential Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

09 Aug 2025, 01:01 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:

Dear Area Chairs,

We would like to respectfully clarify several points in Reviewer 2wfr's latest comment that appear to be factually inaccurate or inconsistent with the record. While we fully respect the reviewer's right to assess novelty and significance, it is important that these judgments are made on the basis of an accurate understanding of the work as submitted and revised. Our intent here is not to dispute subjective assessments but to ensure that the final decision is informed by correct information.

First, the reviewer states that no new experiments were conducted during the rebuttal phase. This is not correct. In response to reviewer feedback, we ran and reported four new imbalanced-data evaluations at 1% and 10% phishing prevalence using two random seeds for statistical robustness, few-shot variation tests with k-values of 1, 3, and 9 to validate our k = 6 choice, and latency profiling across approximately 1,000 queries per model to evaluate throughput and cost trade-offs. These results were described in the rebuttal and integrated into the revised appendix, thereby expanding the experimental scope beyond the original submission.

Second, the review misrepresents the scale of the evaluation, describing it as involving "1000 questions." The actual workload was substantially larger: 10,000 URLs evaluated under two prompting modes across three models, resulting in 60,000 total API calls. Each output was processed across six different evaluation metrics. This level of analysis required significant infrastructure for batching, managing API rate limits, and processing large result sets, and cannot be accurately described as simply "calculating sums."

Third, while the reviewer acknowledges that we explained differences from Nasution et al. (2025), their summary omits key distinctions. Nasution's study used a small, balanced dataset of 1,000 URLs and evaluated only open-source models with full weight access, enabling fine-tuning and architecture-specific optimizations. It did not consider latency, API constraints, class imbalance, or deployment-related performance. Our work, by contrast, evaluates a tenfold larger dataset of 10,000 URLs using proprietary, closed-weight APIs under real operational constraints, handles heterogeneous API input requirements, and tests robustness under multiple class-imbalance scenarios. We also report a more comprehensive metric suite including Accuracy, Precision, Recall, F1, AUROC, and AUPRC. These differences reflect distinct research questions and application contexts—ours focusing on the real-world operational performance of widely used commercial models.

Fourth, the reviewer characterizes our figures as "illegible." All figures in our submission are vector-based PDFs using 11–12 pt fonts at 300 DPI resolution. When viewed at appropriate zoom levels, the text remains crisp and fully readable without pixelation. While legibility can vary depending on PDF viewer scaling settings, the figures meet or exceed typical conference publication standards.

Fifth, on writing quality, the reviewer describes the prose as "poor" or "LLM-generated" but explicitly states this is subjective.

Sixth, the reviewer questions our use of a GPU, assuming the workload involved only ~1,000 queries. As noted above, the experiment consisted of 60,000 API calls. We used a Colab Pro+ A100 instance not for model inference, but for its faster CPU, large RAM, and stable environment for managing large-scale API workflows, latency logging, and post-processing of results. A high-CPU machine would have sufficed for inference, but the GPU-enabled instance provided the needed computational headroom.

Finally, the reviewer is the only one to claim that the paper "does not use the correct submission template." Our LaTeX source was prepared directly with the official NeurIPS style file for initial submissions. The compiled PDF conforms to the conference's formatting requirements for margins, font size, line spacing, and section structure. No other reviewer raised this concern, and the submission system accepted the manuscript without any formatting warnings. If there is a specific deviation the reviewer has in mind, we are happy to address it in the camera-ready version, but we believe the current manuscript adheres to the official template.

We respect that the reviewer maintains a "Strong Reject" recommendation based on their perception of novelty and scope. However, several of their critiques are based on factual misunderstandings or omissions. We believe the reviewer will discontinue to mischaracterize our work during this discussion phase.

If this paper is rejected, NeurIPS will miss publishing the only reproducible, standardized benchmark of frontier LLMs for phishing URL detection, leaving industry, defense, and corporate sectors without an authoritative reference point for model selection in one of the fastest-growing cyber threat domains.

Thank you for your attention to this matter.

Replying to Official Comment by Authors

Official Comment by Reviewer 2wfr

Official Comment by Reviewer 2wfr 09 Aug 2025, 06:59

Program Chairs, Reviewer 2wfr, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

1. I understand the differences, however I still see this work as mostly an extension to Nasution et al. (2025), without sufficient novelty to be considered a standalone conference submission.
2. Apologies for the imprecise wording, I meant experiments expanding substantially expanding the scope of this work or baselining with other methods, which were my concerns.

Regardless of the technical details of figures, I do believe a good figure should be mostly readable without relying on the reader zooming in.

In my understanding the paper uses the accepted version of the submission, while submissions should be using the 'submission' version, with enumerated lines on the sides and the bottom saying 'Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.'. All other papers in my stack are also using this format.

Official Review of Submission23634 by Reviewer fjVz

Official Review by Reviewer fjVz 28 Jun 2025, 05:38 (modified: 18 Sept 2025, 13:00)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer fjVz Revisions (/revisions?id=6czONVsiTh)

Summary:
The paper assesses the effectiveness of LLMs in detecting phishing URLs. The authors benchmark three models, GPT-4o, Claude-3.7-sonnet, and Grok-3-Beta, using zero-shot and few-shot prompting methods. The authors conclude that few-shot prompting improves performance across multiple LLMs.

Strengths And Weaknesses:
The paper presents a clearly written but technically unremarkable evaluation of three proprietary large language models on the task of phishing URL detection using zero-shot and few-shot prompting. While the writing is organised and the methods are described with sufficient clarity to allow replication, the actual content lacks depth and rigour. The experimental setup is minimal, and the evaluation focuses solely on surface-level metrics without any substantive analysis of failure cases, robustness, or real-world applicability. No effort is made to compare the results against established phishing detection systems.

The claims made by the authors are modest, yet even these are poorly supported by the presented data. The choice of models is arbitrary, the task formulation is simplistic, and the results are incremental improvements in standard metrics. In terms of significance, the work does not address a meaningful research gap, nor is there any indication that its findings will have an impact on either the academic community or practical applications.

In terms of originality, the contribution is negligible. The idea of using LLMs with zero- and few-shot prompting for URL classification has already been explored in prior work, e.g., Rashid, Fariza, Nishavi Ranaweera, Ben Doyle, and Suranga Seneviratne. "LLMs are one-shot URL classifiers and explainers." Computer Networks 258 (2025): 111004. Nasution, Arbi Haza, Winda Monika, Aytug Onan, and Yohei Murakami. "Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering." Information 16, no. 5 (2025): 366. The present submission fails to go beyond this prior work in any substantive way. It neither refines the task, introduces novel methods, nor offers new theoretical or empirical understanding.

Quality: 1: poor
Clarity: 4: excellent
Significance: 1: poor
Originality: 1: poor
Questions:
None

Limitations:
Yes

Rating: 1: Strong Reject: For instance, a paper with well-known results or unaddressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
Having read the other reviews as well, I confirm my assessment that the paper should not be accepted.

Rebuttal by Authors

Rebuttal by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

29 Jul 2025, 18:42 (modified: 31 Jul 2025, 16:53) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=cXeBoHdRiB)

Rebuttal:

Thank you for your detailed review.

[Weakness 1]

Contrary to the characterization of "minimal experimental setup," our evaluation encompasses systematic analysis across multiple dimensions. The original paper presents 6 comprehensive experiments evaluating GPT-4o, Claude-3.7, and Grok-3-Beta using balanced datasets with rigorous metrics including Accuracy, Precision, Recall, F1, AUROC, and AUPRC.

To address concerns about statistical robustness, we tested additional experiments using imbalanced datasets with multiple seeds [123, 456] across realistic 1% and 10% phishing ratios : GPT-4o: 92.0% ± 1.4% accuracy Claude: 89.2% ± 1.2% accuracy Grok: 96.2% ± 1.3% accuracy

The claim of "lacks depth and rigour" overlooks our comprehensive methodological approach. Our statistical validation shows remarkably consistent performance patterns. We tested on the imbalanced dataset that models demonstrate exceptional stability: GPT-4o: ±1.4% variance across conditions Claude: ±1.2% variance across conditions Grok: ±1.3% variance across conditions This proves our sampling methodology captures underlying data characteristics rather than artifacts. The computational analysis reveals distinct trade-offs measured locally: GPT-4o: 0.57s average latency (fastest inference) Claude: 1.09s latency (optimal recall for threat detection) Grok: 0.68s latency (superior accuracy)

Regarding "real-world applicability" concerns, our approach directly addresses practical deployment scenarios. We tested locally on imbalanced datasets reflecting real-world conditions where phishing URLs constitute <1% of traffic: At 1% phishing ratio (10 phishing + 990 legitimate URLs): Grok: 97.0% accuracy GPT-4o: 92.6% accuracy Claude: 89.2% accuracy At 10% phishing ratio (100 phishing + 900 legitimate URLs): Grok: 95.4% accuracy GPT-4o: 91.5% accuracy Claude: 89.1% accuracy

The assertion about "surface-level metrics" mischaracterizes our analysis depth. Beyond standard accuracy measures, our evaluation provides behavioral insights through confusion matrix analysis. We found locally distinct error patterns using 1% phishing ratio datasets: GPT-4o: 8.2% false positive rate (81/990 legitimate URLs misclassified) Claude: 11.9% false positive rate (118/990 legitimate URLs misclassified) Grok: 3.5% false positive rate (35/990 legitimate URLs misclassified) Concerning "no comparison against established systems," our systematic comparison of proprietary models addresses a critical practical gap. We tested locally with k-value optimization across k=[1,3,9] using 10% phishing ratio datasets: GPT-4o: improves from 83.3% (k=1) to 94.2% (k=9) Claude: optimal at k=9 with 95.1% accuracy Grok: peak performance at k=1 with 96.9% accuracy

Our evaluation provides practitioners with evidence-based guidance for selecting among commercial APIs, quantifies operational trade-offs through comprehensive analysis, and validates performance under realistic deployment conditions.

[Weakness 2]

In response to "Claims Are Poorly Supported": The paper demonstrates substantial performance differences: Grok-3-Beta achieved 94.05% accuracy, Claude-3.7 reached 92.50%, and GPT-4o attained 90.50% in few-shot settings. These 1.55-3.55 percentage point gaps represent significant operational differences, not incremental improvements. Additional local testing on imbalanced datasets with seeds [123, 456] confirmed consistent model rankings: Grok: 96.2% ± 1.3% Claude: 89.2% ± 1.2% GPT-4o: 92.0% ± 1.4% Standard deviations below 1.4% demonstrate statistical robustness across multiple experimental conditions.

In response to "Choice of Models Is Arbitrary": GPT-4o, Claude-3.7, and Grok-3-Beta represent the three dominant commercial APIs available to enterprise security teams. Organizations must select between these specific commercial APIs for production systems. Academic comparisons using unavailable models provide no actionable guidance for practitioners making actual deployment decisions. Addressing "Task Formulation Is Simplistic": Binary classification reflects production constraints where enterprise security systems require deterministic 0/1 decisions within milliseconds. The 10-token response limit and temperature=0 settings mirror actual API deployment requirements, not oversimplification.

In response to "Results Are Incremental Improvements": Local testing on 1% phishing ratio datasets (10 phishing + 990 legitimate URLs) revealed substantial operational differences: Grok: 3.5% false positive rate (35/990 misclassified) GPT-4o: 8.2% false positive rate (81/990 misclassified) Claude: 11.9% false positive rate (118/990 misclassified) Organizations processing 1 million URLs daily experience 35,000 versus 119,000 false alarms - a 240% difference directly impacting analyst workload, investigation costs, and system reliability.

In response to "Does Not Address Meaningful Research Gap": Prior studies by Rashid et al. and Nasution et al. evaluated different model combinations, datasets, or open-source alternatives. No existing work provides systematic benchmarking of these three specific proprietary APIs under identical conditions. This gap directly affects enterprise security teams unable to make evidence-based selections between their actual deployment options.

In response to "No Practical Impact": Local testing measured distinct computational profiles enabling deployment optimization: GPT-4o: 0.57s average latency (fastest for real-time systems) Claude: 1.09s latency Grok: 0.68s latency (balanced performance) K-value optimization testing on 10% phishing ratio datasets revealed model-specific strategies: GPT-4o: improves from 83.3% (k=1) to 94.2% (k=9) - benefits from more examples Claude: optimal at k=9 with 95.1% - requires maximum context Grok: peak at k=1 with 96.9% - performs best with minimal examples This systematic benchmarking enables evidence-based selection between the three major commercial APIs, addressing a critical practical need in enterprise cybersecurity deployment.

[Weakness 3]

Originality and Contribution: The studies cited, Rashid et al. (2025) and Nasution et al. (2025), investigate similar tasks but differ meaningfully from our work in terms of goals, evaluation design, and practical relevance.

Focus on Commercial APIs, Not Open Models: Nasution et al. tested 21 open-source models like LLaMa3 and Mistral. Our paper evaluates three widely used commercial APIs: GPT-4o, Claude-3.7-sonnet, and Grok-3-Beta, all under the same classification task. These models are accessed through public APIs with distinct input formats and architectural constraints. That practical deployment aspect is not explored in the cited works. Balanced Dataset, Unified Task Setup: Our submitted paper uses a balanced dataset of 10,000 URLs (5,000 phishing and 5,000 legitimate) and evaluates each model using six common classification metrics: accuracy, precision, recall, F1, AUROC, and AUPRC. In contrast, Rashid et al. used smaller samples of around 1,000 URLs and focused more on interpretability through one-shot examples rather than consistent performance benchmarking.

Model-Specific Prompt Formatting: We constructed prompts according to each model's input style. GPT-4o and Grok use a structured format with system and user messages. Claude requires a single flat text prompt. Prior studies used the same format across all models, which overlooks differences in how models are designed to process input.

Few-shot Behavior Varies by Model: In our paper, we tested both zero-shot and few-shot settings using six-shot prompts (3 phishing and 3 legitimate examples). Few-shot prompting improved all models. For example, Grok-3-Beta increased from 89.1 percent to 94.0 percent accuracy. Claude-3.7-sonnet increased from 87.6 percent to 92.5 percent. These kinds of improvements were not studied in the prior work.

Additional Local Testing for Imbalance and Latency (Not in Paper): We also tested on four imbalance datasets ratios of 1 percent and 10 percent, across different random seeds (123 and 456). On a 1 percent phishing test set (10 phishing and 990 legitimate), Grok reached 97.6 percent accuracy with a 3.5 percent false positive rate. Claude dropped to 89.2 percent accuracy with 11.9 percent false positives. These results were produced after submission and are shared only to show how our setup adapts to more realistic traffic patterns. The cited studies did not explore imbalanced scenarios.

We also measured how long each model took to process a single URL. GPT-4o averaged 0.57 seconds, Claude 1.15 seconds, and Grok 0.68 seconds. These differences are important for real-world deployments that require fast response times. Latency data like this was not reported in Rashid et al. or Nasution et al.

Our paper introduces a side-by-side comparison of widely used commercial language models for phishing URL detection, using a consistent evaluation setup, real-world API constraints, and metrics that matter for deployment. Local tests run after submission further show how these models behave under class imbalance and time pressure. These contributions address areas not covered in prior work.

Replying to Rebuttal by Authors

Official Comment by Reviewer fjVz

Official Comment by Reviewer fjVz 04 Aug 2025, 11:58 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Although I did not have any questions for the authors, I acknowledge their response and thank them for taking the time to consider my review.

I noted the authors' effort, but the following points remain: 1. The study employs three proprietary large language models (LLMs) for a binary classification task on an already published dataset of phishing URLs. 2. No traditional machine learning benchmarks are provided for comparison. 3. The analysis is limited to basic performance metrics. For instance, the authors do not consider the monetary cost associated with classification, thereby constraining the assessment of real-world applicability. In their rebuttal, the authors themselves highlight time-to-classification as a critical factor when critiquing other work. Yet, a latency of 0.5 seconds per URL (at best) is relatively high for practical deployment. This is consistent with their own comment to reviewer fjVz, where they state: "Binary classification reflects production constraints where enterprise security systems require deterministic 0/1 decisions within milliseconds." 4. The central claim is limited in scope (given the existing literature on the benefits of few-shot prompting and other research on LLMs and phishing URLs) and weakly substantiated (the study is based on a single dataset, three proprietary LLMs that required different prompting strategies, and no comparison with other ML approaches).

I acknowledge the authors' rebuttal, but I do not intend to revise my evaluation.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer fjVz

Mandatory Acknowledgement by Reviewer fjVz 04 Aug 2025, 12:00 (modified: 12 Aug 2025, 02:48)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions (/revisions?id=5Hnyi73NwK)

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Official Comment by Authors

Official Comment
by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))
09 Aug 2025, 01:38 (modified: 09 Aug 2025, 01:56) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=nWJSngQI3B)

Comment:
Thank you for your response.

Our work provides the first direct performance comparison of the three major commercial LLMs available via public APIs for cybersecurity applications. Commercial models require different evaluation approaches due to API constraints, making our standardized framework methodologically necessary for fair comparison.

Regarding originality, we agree that zero/few-shot LLM use for URL detection is not new, and we cite relevant works including Nasution et al. (21 open-source models, balanced small set, CoT) in our manuscript. However, our contribution is distinct: the first reproducible benchmark of GPT-4o, Claude-3.7, and Grok-3-Beta on a shared phishing URL task, with proprietary API constraints, latency/cost analysis, and robustness under imbalance. This fills a gap in the literature rather than duplicating prior work. We have framed our contribution statement to reflect this and explicitly contrasted our results with prior benchmarks (e.g., Grok few-shot ~94% F1 vs. Llama3-70B's 91.2% in Nasution) in the rebuttal.

Our results reveal significant practical differences: Claude-3.7 achieves 95.26% recall versus Grok-3-Beta's 94.92% precision, informing deployment trade-offs between threat detection and false alarm reduction. We discuss misclassification patterns, for example, Grok-3-Beta produces fewer false positives (high precision) while Claude-3.7 has fewer false negatives (high recall), a distinction evident in the confusion matrices. We interpret this as Grok being more conservative and Claude more aggressive in flagging phishing, which has deployment implications. On practical impact, while the few-shot accuracy spread between the best and worst model is ~3.5 percentage points, in large-scale deployments this difference translates to tens of thousands of incidents. At 1% prevalence, Claude's false positive rate would generate ~119,000 false alerts per million URLs versus Grok's ~35,000—an operationally critical difference. Academically, showing that few-shot prompting consistently yields 2–5 point F1 improvements confirms the utility of prompt engineering in this domain and suggests research directions into why Grok excels.

We selected GPT-4o, Claude-3.7, and Grok-3 deliberately, as they are leading commercial services in 2025 and represent real-world procurement choices. The binary classification task is intentionally simple to reflect operational phishing filters that require a deterministic allow/block decision. Restricting outputs to "0" or "1" and using temperature = 0 avoids variability and supports reproducibility; richer tasks can be explored in future work but would introduce additional confounds.

Classical and deep learning baselines have already been extensively benchmarked on malicious URL datasets, including PhiUSIIL, in prior studies. Our work complements these efforts by analyzing how modern proprietary LLMs perform in zero-shot and few-shot settings without retraining or internal access, a realistic deployment scenario not covered in earlier research.

While our original focus was latency as a proxy, we now estimate API costs: Based on current API pricing, GPT-4o costs approximately $0.004 per URL classification, Claude Sonnet 3.7 costs $0.015, and Grok-3 costs $0.006, making cost-performance evaluation critical for enterprise adoption. For enterprise deployments, such costs may be acceptable given the performance benefits.

Our latency measurements reveal significant differences: GPT-4o (0.56–0.60s), grok-3 (0.59–0.77s), and Claude Sonnet 3.7 (1.09–1.30s) per classification. While these latencies are higher than traditional ML methods, they reflect the current state of commercial LLM APIs. Our analysis provides essential baseline data for organizations evaluating whether the enhanced detection capabilities justify the increased latency and cost compared to classical approaches. The "milliseconds" requirement mentioned in our rebuttal refers to the final binary decision output, not the entire classification pipeline. In practice, URL screening often occurs in batch processes or during threat intelligence analysis where sub-second latencies are acceptable. However, we acknowledge that for real-time web filtering applications, current LLM latencies may limit deployment scenarios.

Our evaluation provides practitioners with concrete performance and operational trade-offs: Claude offers superior recall (95.26%) but highest latency and cost, while Grok provides optimal precision (94.92%) with moderate speed and lowest cost. This enables informed decisions about which commercial LLM best fits specific security workflows and budget constraints.

The paper does not aim to establish few-shot prompting as novel. It examines how detection performance varies across proprietary LLMs when applied to phishing URLs under consistent deployment constraints. The focus is not general prompting behavior, but model-specific differences in a fixed classification task.

Replying to Official Comment by Authors

Author AC Confidential Comment by Authors

Author AC Confidential Comment

by Authors (Prashanth BusiReddyGari (/profile?id=~Prashanth_BusiReddyGari1), Najmul Hasan (/profile?id=~Najmul_Hasan1))

09 Aug 2025, 01:48 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:

Dear Area Chairs,

Our work provides the first direct, reproducible benchmark of three major proprietary LLMs (GPT-4o, Claude-3.7-Sonnet, and Grok-3-Beta) on phishing URL detection via public APIs. This evaluation requires a standardized framework because these APIs have different input structures and constraints, which makes direct comparison non-trivial. Algorithm 1 in our paper encodes these adaptations, ensuring identical task semantics while accommodating each model's API. This methodology enables fair, deployment-relevant comparisons not previously available in the literature.

Reviewer fjVz cites Rashid et al. (2025) and Nasution et al. (2025) to suggest our work duplicates prior studies. While inspired by these works, our scope differs. Nasution tests 21 open-source models on 1,000 URLs with full weight access, no latency or imbalance analysis, and homogeneous API structures. Our study tests three closed-source, enterprise-grade APIs on 10,000 URLs per mode, with evaluations across balanced and imbalanced datasets, latency profiling, and prompt format adaptation. To our knowledge, no prior study has benchmarked these three commercial LLMs side-by-side in this way.

Classical phishing detectors and deep learning methods have been extensively evaluated on PhiUSIIL and related datasets (e.g., Rafsanjani et al., Zhou et al.), achieving ~95–97% accuracy with full supervision. Our aim was to complement these works by focusing on training-free zero/few-shot LLM use, which is the realistic scenario when deploying proprietary APIs.

We have added explicit API cost estimates to our analysis in our rebuttal. The "milliseconds" reference in our rebuttal pertains to the decision step within enterprise security workflows, not the full inference pipeline. In practice, many deployments batch-process URLs, making sub-second per-item latency acceptable in non-interactive contexts.

The binary classification framing is deliberate. Many security systems ultimately require a deterministic allow/block decision on a URL, without explanation generation. We therefore fixed the output to a single binary label, kept prompts minimal, and set temperature to zero to mirror production constraints. This simplicity enables clean reproducibility and direct cross-model comparison. Our results also show that few-shot prompting consistently improves F1 by 2–5 points, even under these constraints.

While Claude supports system prompts, we used a single-message format to maintain consistent semantic framing across models. GPT-4o and Grok require system-user separation per API guidelines, so our framework adapts format while preserving identical instructions. This avoids prompt architecture as a confounding factor.

In summary, we believe our research study fills a genuine gap: a controlled, reproducible, and deployment-relevant evaluation of proprietary LLMs for phishing URL detection. The additional robustness experiments, latency/cost analysis, and contextualization with classical baselines strengthen both the academic and practical value of the paper. We hope this clarifies the record and provides a more accurate basis for deliberation.

Our work has high practical readability for a wide audience and will provide industry, defense, and corporate security with a vendor-neutral, reproducible benchmark to guide LLM selection, adapt detection to evolving threats without retraining, and support evidence-based procurement and policy decisions.

Thank you.