LLMs para Detecção de Phishing em E-mails: Comparação Controlada com Baselines TF-IDF e Red Flags Auditáveis

Bernardo Dorneles[1], Silvio Quincozes[1,2]

[1,1]AI Horizon Labs – Universidade Federal do Pampa (UNIPAMPA)
[2]PPGCO – Universidade Federal de Uberlândia (UFU)

{bernardodorneles.aluno,silvioquincozes}@unipampa.edu.br

Abstract. *This paper evaluates large language models (LLMs) for phishing detection in e-mails from a cybersecurity-oriented reproducible protocol. We compare four recent LLMs on the same 90-message evaluation set, separated from a 10-message technical calibration step, and contrast them with TF-IDF baselines trained without leakage from the evaluation messages. Qwen3 32B achieved the best LLM performance and perfect phishing recall, while Llama 3.3 70B showed the most balanced precision-recall profile. However, TF-IDF baselines obtained a higher raw F1-score in this textual setting. The contribution is therefore not a claim that LLMs are universally better classifiers, but an empirical analysis of when they add operational value through auditable red flags and structured justifications.*

Resumo. *Este trabalho avalia Grandes Modelos de Linguagem, do inglês, Large Language Models (LLMs) na detecção de phishing em e-mails sob um protocolo reprodutível, controlado e orientado à cibersegurança. Comparamos quatro LLMs recentes em um mesmo conjunto avaliativo de 90 mensagens, separado de uma etapa técnica de calibração com 10 mensagens, e confrontamos os resultados com baselines TF-IDF treinadas sem vazamento do conjunto de avaliação. Entre os LLMs, o Qwen3 32B obteve o melhor desempenho e revocação perfeita para phishing, enquanto o Llama 3.3 70B apresentou o perfil mais equilibrado entre precisão e revocação. Entretanto, as baselines TF-IDF alcançaram maior F1-score bruto neste recorte textual. A contribuição, portanto, não é afirmar superioridade geral dos LLMs, mas analisar quando eles agregam valor operacional por meio de red flags auditáveis e justificativas estruturadas.*

1. Introdução

Phishing combina engenharia social e falsificação de identidade para induzir ações indevidas e permanece ativo em escala elevada (Anti-Phishing Working Group 2025). O problema operacional não é apenas identificar mensagens maliciosas, mas reduzir falsos negativos sem gerar falsos positivos que inviabilizem a triagem humana. Campanhas contemporâneas também exploram SMS (*smishing*), chamadas de voz (*vishing*) e códigos QR (*quishing*) (Anti-Phishing Working Group 2025, Sharevski et al. 2022); este estudo, porém, restringe-se a e-mails textuais. Grandes Modelos de Linguagem, do inglês *Large Language Models* (LLMs), são candidatos atraentes para esse cenário por

combinarem interpretação textual, classificação e geração de justificativas em linguagem natural. Estudos recentes discutem seu uso tanto na criação quanto na detecção de phishing (Heiding et al. 2024, Nahmias et al. 2024, Li et al. 2024, Bustamante et al. 2025). Ainda assim, uma avaliação útil para segurança precisa evitar três armadilhas metodológicas: comparar modelos em amostras diferentes, confundir calibração de *prompt* com avaliação final e tratar justificativas geradas automaticamente como explicações necessariamente corretas.

Este trabalho investiga essa lacuna por meio de uma comparação controlada entre GPT-OSS 120B, Llama 3.3 70B, Qwen3 32B e Compound em 90 e-mails avaliativos, separados de uma etapa prévia de calibração técnica com 10 e-mails. Os mesmos 90 e-mails também são utilizados para avaliar *baselines* baseados em *Term Frequency–Inverse Document Frequency* (TF-IDF), combinados com Regressão Logística e *Random Forest*, treinados sem vazamento em relação ao protocolo principal. A pergunta de pesquisa que orienta o estudo é: em um protocolo controlado, que abordagem oferece melhor compromisso entre detecção de phishing, preservação de e-mails legítimos e produção de sinais auditáveis?

A principal contribuição deste trabalho é uma avaliação empírica e incremental do compromisso entre desempenho preditivo e auditabilidade na triagem de *phishing*. O protocolo compara, nas mesmas instâncias e sem vazamento, quatro LLMs e duas *baselines* TF-IDF, separando calibração e avaliação e examinando falsos positivos, falsos negativos e *red flags*. Assim, o estudo identifica quando saídas estruturadas podem agregar valor operacional, mesmo sem superar métodos clássicos em *F1-score*.

2. Fundamentação Teórica

O phishing por e-mail constitui um problema sociotécnico, pois sua efetividade não depende apenas de elementos técnicos da mensagem, mas também da forma como linguagem, contexto, urgência, autoridade simulada e confiança do usuário são explorados para induzir uma ação indesejada (Jakobsson and Myers 2006, Sheng et al. 2010). Dessa forma, a avaliação de mecanismos defensivos não deve se restringir à acurácia média. Em cenários operacionais, é igualmente importante observar o tipo de erro produzido: falsos negativos podem expor usuários e organizações a comprometimentos, enquanto falsos positivos podem ampliar a carga de revisão manual, gerar bloqueios indevidos e reduzir a confiança nos controles de segurança.

Métodos clássicos baseados em bolsa de palavras e TF-IDF permanecem relevantes nesse contexto porque regularidades lexicais ainda permitem distinguir parte significativa de mensagens maliciosas e legítimas (Salton and Buckley 1988, Fette et al. 2007, Bergholz et al. 2010). Embora tais métodos não capturem plenamente aspectos pragmáticos, contextuais ou persuasivos do phishing, eles oferecem vantagens importantes: baixo custo computacional, reprodutibilidade, facilidade de implantação e comportamento relativamente previsível. Neste trabalho, portanto, essas técnicas não são tratadas como *baselines* fracas ou meramente ilustrativas, mas como alternativas leves e operacionalmente plausíveis para *gateways*, filtros corporativos e *pipelines* de triagem.

LLMs ampliam esse espaço de investigação ao combinar interpretação textual, classificação e geração de justificativas em linguagem natural. Essa característica os torna especialmente atrativos para tarefas em que a decisão automatizada precisa ser acompa-

nhada por sinais compreensíveis para analistas humanos. No entanto, as justificativas produzidas por LLMs devem ser interpretadas com cautela: elas podem ser úteis como artefatos de auditoria e apoio à revisão, mas não devem ser confundidas com explicações causais ou garantias de acesso ao raciocínio interno do modelo.

Neste artigo, a interpretabilidade é tratada de forma pragmática. Métricas como precisão, revocação (*recall)*, especificidade e *F1-score* resumem diferentes dimensões do desempenho, mas não indicam, por si só, quais indícios motivaram uma decisão. Por isso, as *red flags* geradas pelos LLMs são analisadas como sinais estruturados para auditoria, comparação e apoio operacional, e não como explicações definitivas do processo decisório. De modo complementar, a análise SHAP é restrita a um Random Forest auxiliar treinado com essas *flags*, servindo apenas para qualificar a influência relativa dos sinais estruturados na tarefa de triagem (Lundberg and Lee 2017).

3. Trabalhos Relacionados

A literatura sobre phishing por e-mail reúne três eixos principais: fatores humanos, classificação textual e, mais recentemente, mecanismos baseados em LLMs. Estudos clássicos caracterizam phishing como uma forma de imitação persuasiva e abuso de confiança, mostrando que a resposta do usuário depende de fatores como contexto, perfil, familiaridade com a ameaça e percepção de risco (Jakobsson and Myers 2006, Sheng et al. 2010). Essa base sociotécnica justifica a análise de sinais recorrentes em mensagens maliciosas, como urgência, autoridade simulada, saudação genérica, inconsistências de identidade e solicitação de credenciais ou ações sensíveis.

No campo da detecção textual, trabalhos anteriores demonstram que atributos lexicais, estruturais e semânticos podem sustentar classificadores supervisionados eficazes para identificar phishing (Fette et al. 2007, Bergholz et al. 2010). Mesmo com o avanço de modelos mais expressivos, resultados recentes com abordagens supervisionadas e híbridas reforçam que métodos não baseados em LLMs continuam competitivos e devem ser incluídos em comparações justas (Mahendru and Pandit 2024, Nair et al. 2025). Essa constatação é central para o presente estudo, pois evita avaliar LLMs de forma isolada, sem um ponto de referência simples, reprodutível e operacionalmente plausível.

LLMs têm sido investigados sob uma perspectiva ambivalente: ao mesmo tempo em que podem apoiar a criação de mensagens convincentes, automatizar variações linguísticas e potencializar campanhas de *spear phishing*, também podem ser empregados na classificação, triagem e justificativa de e-mails suspeitos (Heiding et al. 2024, Nahmias et al. 2024, Koide et al. 2024a, Chataut et al. 2024). Trabalhos correlatos ampliam esse uso para detecção em páginas Web, ataques baseados em URL e cenários multimodais (Li et al. 2024, Koide et al. 2024b). Em contraste, o presente artigo adota um recorte mais controlado: concentra-se em e-mails textuais, aplica os mesmos exemplos avaliativos a todos os modelos e analisa os erros sob uma perspectiva operacional, considerando tanto o risco de falsos negativos quanto o custo de falsos positivos.

A explicabilidade também tem recebido atenção em segurança, especialmente pela necessidade de tornar decisões automatizadas mais úteis para analistas. Diferentes trabalhos exploram explicações textuais, LIME, mecanismos de atenção ou SHAP para apoiar interpretação, depuração e confiança em classificadores (Lim et al. 2025, Uddin and Sarker 2024, Lundberg and Lee 2017). No contexto nacional, estudos recentes

também investigaram a caracterização de phishing com LLMs e *prompts* padronizados (Bustamante et al. 2025).

Em relação aos estudos anteriores, o diferencial não está em cada escolha metodológica isolada, mas na análise conjunta, sobre as mesmas instâncias, do desempenho de LLMs e *baselines* leves e da utilidade operacional das *red flags*. O estudo se posiciona, portanto, como uma avaliação comparativa e incremental, não como a proposição de um novo método.

4. Metodologia

O estudo utiliza 100 mensagens extraídas do dataset público *Phishing Emails Dataset*, divididas em 10 e-mails para calibração técnica e 90 e-mails para avaliação comparativa. A etapa de calibração foi empregada exclusivamente para verificar o *prompt*, o *parser* JSON, os registros de execução e a taxonomia de *red flags*; nenhuma mensagem dessa etapa foi incluída nas métricas principais. Essa separação reduz o risco de ajustar o protocolo ao conjunto avaliativo e preserva a independência entre calibração e teste.

A amostra foi definida para viabilizar a execução completa e rastreável do protocolo em múltiplos modelos. Seu tamanho permite comparações pareadas nas mesmas instâncias, mas não sustenta generalizações amplas sobre outros datasets, idiomas ou contextos organizacionais. O estudo deve, portanto, ser interpretado como uma avaliação controlada e exploratória.

As classes consideradas são `phishing` e `safe`, com `phishing` definido como classe positiva. Os 90 e-mails avaliativos foram selecionados por uma preparação fixa com semente 2026, resultando em 36 mensagens rotuladas como phishing e 54 mensagens rotuladas como seguras. Essa distribuição decorre da amostragem reprodutível com a semente definida, e não de uma imposição manual de balanceamento. Todos os modelos foram executados com temperatura 0, utilizando o mesmo esquema de *prompt* e o mesmo *parser* de saída.

A comparação principal inclui apenas modelos que produziram 90 respostas válidas: `groq-gpt-oss-120b`, `groq-llama-3-3-70b`, `groq-qwen3-32b` e `groq-compound`. Como referência operacional, foram treinados classificadores baseados em TF-IDF com Regressão Logística e *Random Forest*. Esses modelos foram treinados sobre 18.531 e-mails remanescentes do dataset bruto, excluindo todos os 100 e-mails do protocolo, incluindo calibração e avaliação, para evitar vazamento de dados.

Cada LLM retorna uma classificação binária e um conjunto padronizado de *red flags*, incluindo sinais como remetente suspeito, urgência, solicitação de dados sensíveis, links suspeitos e erros gramaticais. Essas *flags* são utilizadas para análise de auditoria e erro, mas não foram validadas por especialistas humanos. Portanto, devem ser interpretadas como sinais estruturados produzidos pelo modelo, que podem ser úteis para revisão e comparação, mas não como explicações necessariamente corretas ou causalmente fiéis ao processo decisório interno dos LLMs. De modo complementar, a análise SHAP é aplicada a um modelo supervisionado auxiliar (Random Forest) treinado com essas *flags*, com o objetivo restrito de estimar a influência relativa desses sinais estruturados, sem pretensão de explicar internamente os LLMs.

As métricas consideradas foram acurácia, precisão, revocação, especificidade,

F1-score, taxa de falsos positivos e taxa de falsos negativos. Como o custo dos erros é assimétrico em segurança, a análise enfatiza especialmente a revocação para phishing e a taxa de falsos negativos. Um e-mail de phishing classificado como legítimo representa risco direto de comprometimento, enquanto um e-mail legítimo classificado como phishing tende a produzir fricção operacional, revisão manual e possíveis bloqueios indevidos. Por isso, a acurácia isolada não é suficiente para sustentar a comparação entre abordagens. Além das métricas descritivas, são utilizados intervalos de confiança por *bootstrap* para *F1-score* e revocação, bem como o teste exato de McNemar para comparações pareadas entre classificadores avaliados sobre as mesmas instâncias.

5. Resultados

A Tabela 1 resume o desempenho dos quatro LLMs e das duas *baselines* clássicas no mesmo conjunto avaliativo de 90 e-mails. O primeiro resultado relevante é que todos os LLMs produziram cobertura completa, isto é, 90/90 respostas válidas. Isso é importante porque evita uma comparação enviesada por falhas de execução, respostas inválidas ou exclusão de casos difíceis. Assim, as diferenças observadas refletem principalmente o comportamento classificatório de cada abordagem no protocolo definido.

Tabela 1. Métricas principais dos LLMs e das *baselines* clássicas no mesmo conjunto avaliativo de 90 e-mails.

Modelo	Tipo	Acurácia	Precisão	Revocação	Especificidade	*F1-score*	FP	FN	Cobertura
TF-IDF + Regressão Logística	*Baseline*	0,9778	0,9474	1,0000	0,9630	0,9730	2	0	90/90
TF-IDF + *Random Forest*	*Baseline*	0,9778	0,9722	0,9722	0,9815	0,9722	1	1	90/90
Qwen3 32B	LLM	0,9333	0,8571	1,0000	0,8889	0,9231	6	0	90/90
Llama 3.3 70B	LLM	0,9222	0,8919	0,9167	0,9259	0,9041	4	3	90/90
Compound	LLM	0,9000	0,9091	0,8333	0,9444	0,8696	3	6	90/90
GPT-OSS 120B	LLM	0,8778	0,8378	0,8611	0,8889	0,8493	6	5	90/90

O principal achado é que as *baselines* baseadas em TF-IDF permaneceram altamente competitivas e superaram os LLMs em desempenho preditivo bruto neste recorte. A Regressão Logística atingiu o melhor *F1-score* geral e não produziu falsos negativos, enquanto o *Random Forest* apresentou desempenho praticamente equivalente, com apenas um falso positivo e um falso negativo. Esse resultado reforça que, para triagem textual de phishing em conjuntos nos quais regularidades lexicais são discriminativas, métodos clássicos ainda constituem referências fortes, leves e operacionalmente plausíveis. Portanto, a contribuição dos LLMs não deve ser interpretada como substituição automática de classificadores tradicionais, mas como uma alternativa com propriedades distintas, sobretudo quando a saída estruturada e a auditabilidade são relevantes.

Entre os LLMs, o Qwen3 32B apresentou o perfil mais favorável para cenários em que a prioridade é reduzir o risco de comprometimento. O modelo obteve revocação perfeita para phishing, eliminando falsos negativos no conjunto avaliativo, mas ao custo de seis falsos positivos. Esse comportamento sugere uma postura mais sensível a indícios de ameaça: operacionalmente, ele tende a preservar a segurança, mas transfere maior carga para revisão humana ao sinalizar e-mails legítimos como suspeitos. Em ambientes nos quais falsos negativos são mais caros que falsos positivos, esse perfil pode ser aceitável; em contextos com alto volume de mensagens e baixa capacidade de triagem manual, o mesmo comportamento pode gerar atrito excessivo.

O Llama 3.3 70B apresentou um compromisso mais equilibrado entre detecção de phishing e preservação de e-mails legítimos. Embora não tenha alcançado a revocação perfeita do Qwen3 32B, produziu menos falsos positivos e manteve desempenho consistente nas diferentes métricas. Esse resultado sugere um comportamento menos agressivo na emissão de alertas, com menor penalização sobre mensagens legítimas. O Compound, por sua vez, foi o mais conservador entre os LLMs: reduziu falsos positivos, mas deixou passar mais mensagens de phishing. Esse perfil pode parecer atraente pela menor fricção operacional, mas é arriscado em segurança, pois falsos negativos representam exposição direta a ataques. O GPT-OSS 120B teve o desempenho mais baixo entre os LLMs avaliados, combinando falsos positivos e falsos negativos em quantidade maior que os modelos mais competitivos, ainda que tenha mantido cobertura completa.

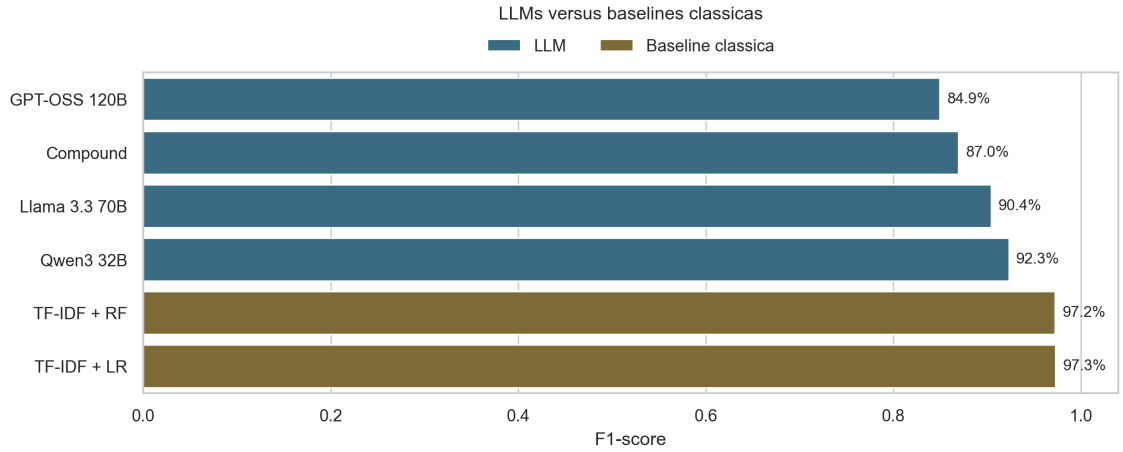

Figura 1. Comparação de *F1-score* entre os quatro LLMs e as duas *baselines* clássicas avaliadas no mesmo conjunto de 90 e-mails.

A Figura 1 evidencia essa separação entre desempenho preditivo bruto e potencial de apoio operacional. As *baselines* ocupam a parte superior do comparativo de *F1-score*, indicando que o simples uso de LLMs não garante ganho de classificação. Ao mesmo tempo, a proximidade do Qwen3 32B em relação às abordagens clássicas sugere que alguns LLMs podem ser competitivos mesmo sem treinamento específico sobre o conjunto completo. A questão central, portanto, passa a ser menos ‘qual modelo acerta mais em média” e mais ‘qual modelo oferece o melhor compromisso entre risco, custo de revisão e utilidade dos sinais produzidos”.

A análise de erros da Figura 2 reforça que modelos com *F1-score* próximos podem ser operacionalmente muito diferentes. O Qwen3 32B e a Regressão Logística, por exemplo, não produziram falsos negativos, mas diferem substancialmente em falsos positivos. Já o Compound reduz alertas indevidos, mas aumenta o risco de liberar mensagens maliciosas. Essa assimetria é decisiva em phishing: falsos positivos afetam produtividade e confiança nos filtros, enquanto falsos negativos podem resultar em roubo de credenciais, instalação de malware ou comprometimento organizacional. Assim, a análise por tipo de erro é mais informativa que a acurácia isolada.

Em relação às *red flags*, os modelos revelaram estilos distintos de sensibilidade. O Qwen3 32B acionou com frequência sinais associados à forma da mensagem, como erros gramaticais, saudação genérica e formatação estranha, além de marcar e-mails

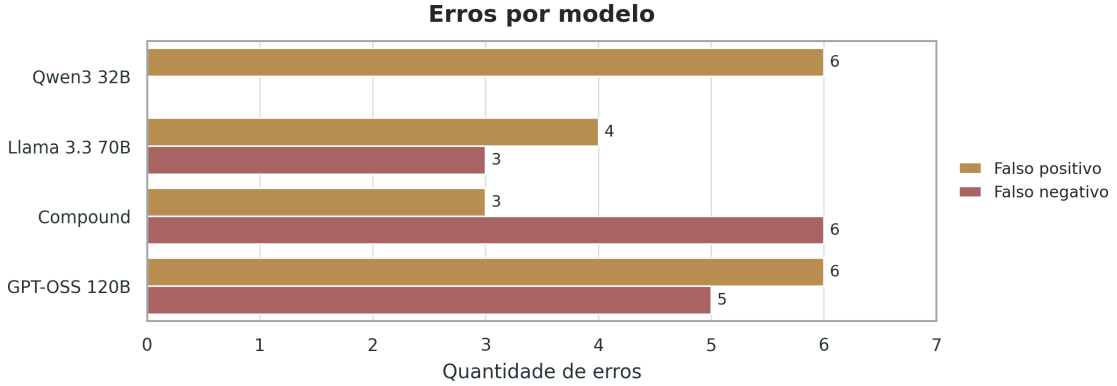

Figura 2. Perfil de erros por modelo. O Compound reduz falsos positivos, mas à custa de mais falsos negativos; o Qwen3 32B elimina falsos negativos, porém com mais alertas indevidos sobre e-mails legítimos.

não solicitados. O Llama 3.3 70B distribuiu suas ativações de modo mais equilibrado entre e-mail não solicitado, remetente suspeito, saudação genérica e domínio suspeito. O Compound produziu menos alertas em geral, coerente com seu perfil mais conservador. Já o GPT-OSS 120B acionou *red flags* com maior intensidade, especialmente erros gramaticais, formatação estranha, e-mail não solicitado e contato ausente ou inconsistente. A Figura 3 sintetiza essas diferenças ao apresentar a frequência média de cada *red flag* por modelo.

Essas diferenças sugerem que as *red flags* não devem ser interpretadas apenas como explicações locais, mas também como indicadores do viés operacional de cada modelo. Modelos que acionam muitos sinais podem favorecer a detecção ampla, mas também tendem a gerar mais falsos positivos quando sinais fracos aparecem em mensagens legítimas. Por outro lado, modelos que exigem evidências mais explícitas podem reduzir alertas indevidos, mas falhar em ataques ambíguos, curtos ou com poucos indícios textuais. Isso aparece especialmente nos falsos negativos do Compound, que em alguns casos envolveram mensagens vazias, incompletas ou com poucos sinais diretos. Nesses casos, a ausência de evidência textual forte parece ter favorecido uma decisão conservadora, ainda que incorreta do ponto de vista de segurança.

A Figura 4 detalha a matriz de confusão do Qwen3 32B, o LLM com melhor *F1-score* e revocação no conjunto avaliativo. O resultado sintetiza o compromisso mais importante desse modelo: nenhum phishing foi classificado como seguro, mas seis mensagens legítimas foram sinalizadas como phishing. Em uma implantação prática, esse perfil poderia ser útil como primeira camada de triagem ou como mecanismo de priorização para revisão humana, desde que o custo dos falsos positivos seja administrável.

A Tabela 2 apresenta os intervalos de confiança por *bootstrap* para as duas *baselines* clássicas e para o melhor LLM no conjunto avaliativo. Os intervalos reforçam a vantagem descritiva das *baselines* em *F1-score*, mas também mostram que a incerteza ainda é relevante, especialmente devido ao tamanho limitado do conjunto avaliativo. O Qwen3 32B mantém revocação perfeita nas reamostragens, mas seu intervalo de *F1-score* é mais amplo, refletindo maior sensibilidade aos falsos positivos.

Nas comparações pareadas com o Qwen3 32B, os testes exatos de McNemar frente à Regressão Logística e ao *Random Forest* resultaram em $p = 0{,}2891$ em ambos os

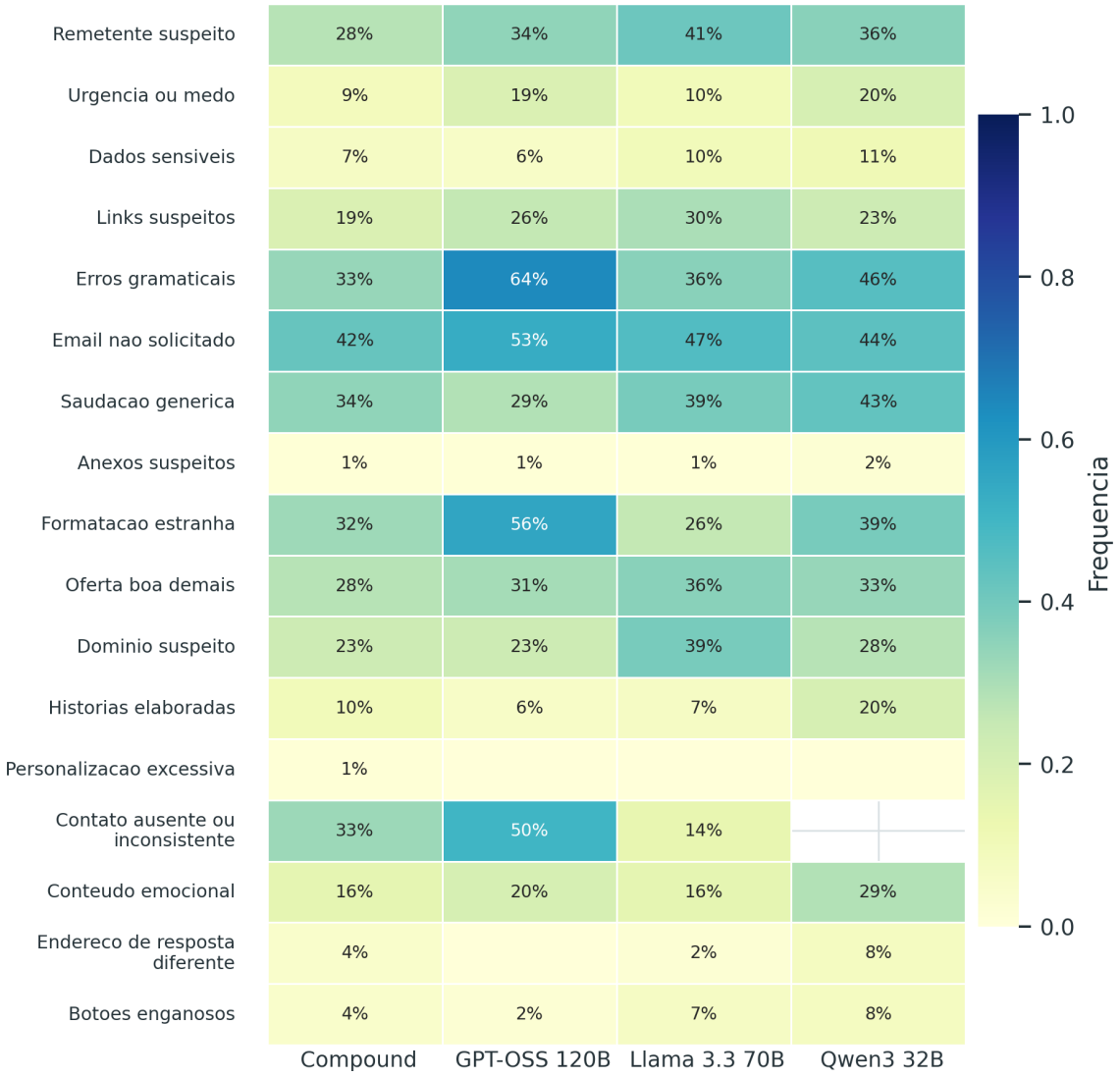

Figura 3. Frequência média de *red flags* por modelo no conjunto avaliativo.

casos. Esse resultado indica que, embora as *baselines* tenham obtido melhor desempenho descritivo no conjunto avaliado, não houve evidência estatística suficiente para rejeitar a hipótese nula ao nível de 5%. Essa constatação não deve ser interpretada como equivalência entre classificadores. Com apenas 90 e-mails, o poder estatístico é limitado, e diferenças operacionalmente relevantes podem não alcançar significância formal.

6. Discussão

Os resultados indicam que a escolha do modelo depende do risco dominante. O Qwen3 32B não deixou passar nenhum phishing no conjunto avaliativo, característica atraente quando falsos negativos são o erro mais crítico. Seu custo foi produzir seis falsos positivos, o que pode aumentar a revisão manual. O Llama 3.3 70B não alcançou revocação perfeita, mas apresentou melhor equilíbrio entre precisão e revocação. O Compound preservou mais e-mails legítimos, com menor taxa de falsos positivos, mas perdeu mais ataques.

As baselines TF-IDF são o principal freio contra uma interpretação exagerada dos resultados. Elas superaram os LLMs em F1-score bruto neste recorte, mostrando que

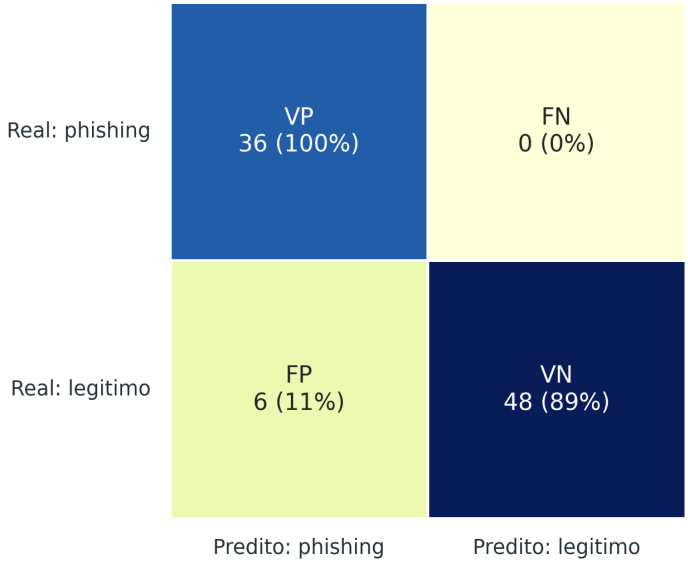

Figura 4. Matriz de confusão do Qwen3 32B.

Tabela 2. Intervalos de confiança por *bootstrap* no conjunto avaliativo de 90 e-mails.

Modelo	*F1-score* médio	IC95% do *F1-score*	IC95% da revocação
TF-IDF + Regressão Logística	0,9727	(0,9296, 1,0000)	(1,0000, 1,0000)
TF-IDF + *Random Forest*	0,9721	(0,9254, 1,0000)	(0,9032, 1,0000)
Qwen3 32B	0,9227	(0,8529, 0,9767)	(1,0000, 1,0000)

métodos leves continuam fortes para e-mails textuais com regularidades lexicais. Assim, o argumento a favor dos LLMs não é o desempenho preditivo universalmente superior, mas a possibilidade de combinar classificação com red flags, justificativas estruturadas e análise operacional de erros.

As *red flags* cumprem papel intermediário de explicabilidade. Elas permitem inspecionar quais sinais aparecem com maior frequência e como se associam a falsos positivos e falsos negativos, aproximando o sistema de uma triagem auditável para analistas (Lim et al. 2025). Ainda assim, não devem ser tratadas como evidência factual automaticamente correta nem como janela fiel para o raciocínio interno do LLM. A análise SHAP reforça essa cautela: as variáveis mais influentes no Random Forest auxiliar foram oferta boa demais (0,1817), remetente suspeito (0,0744), formatação estranha (0,0675), domínio suspeito (0,0337) e conteúdo emocional (0,0312), mas esses valores explicam o modelo auxiliar treinado sobre *red flags*, não os LLMs comparados.

Como contribuição incremental, o valor científico está no desenho experimental controlado, na reprodução dos resultados e na leitura honesta dos limites. O estudo não resolve o problema de detecção de phishing, mas oferece um protocolo incremental e auditável que pode ser ampliado com mais amostras, validação humana das flags e baselines supervisionadas modernas.

7. Disponibilidade dos Artefatos e Reprodutibilidade

O dataset original utilizado neste estudo está disponível publicamente na plataforma Kaggle como "Phishing Emails Dataset" (Subhajournal 2023).[1]

Para preservar a revisão duplo-cega, os artefatos de reprodução não incluem identificação autoral nesta versão. Resultados consolidados, métricas por modelo, baselines, frequência de red flags, figuras, intervalos de confiança, testes de McNemar e scripts auxiliares serão disponibilizados publicamente após a revisão; se exigido pela organização, podem ser fornecidos por repositório anônimo durante a avaliação.

8. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma comparação controlada de quatro LLMs para detecção de phishing em e-mails, com separação explícita entre calibração técnica e avaliação comparativa. No conjunto de 90 e-mails, o Qwen3 32B obteve o melhor F1-score entre os LLMs e revocação perfeita para phishing; o Llama 3.3 70B apresentou o perfil mais equilibrado; o Compound reduziu falsos positivos ao custo de mais falsos negativos; e o GPT-OSS 120B teve desempenho inferior aos demais LLMs, embora com cobertura completa.

Quando comparados às baselines, os LLMs não foram superiores em F1-score bruto: TF-IDF com Regressão Logística e TF-IDF com Random Forest alcançaram os melhores resultados neste recorte textual. Esse achado é relevante para cibersegurança porque evita substituir soluções leves por LLMs sem evidência operacional. O valor potencial dos LLMs aparece principalmente na geração de red flags e justificativas estruturadas que podem apoiar auditoria, triagem e análise de falhas.

Como trabalhos futuros, pretendemos ampliar a amostra, avaliar diferentes sementes, incluir e-mails em português e cenários multilíngues, validar red flags com especialistas humanos, medir custo e latência, testar mensagens adversariais e comparar com transformers supervisionados especializados. Trabalhos nacionais recentes (Bustamante et al. 2025, Quincozes et al. 2025) indicam espaço para estudos maiores; o próximo passo é combinar escala, reprodutibilidade e análise de erro orientada a risco.

Referências

Anti-Phishing Working Group (2025). Phishing activity trends report: 4th quarter 2024. Technical report, APWG. Released March 19, 2025.

Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paass, G., and Strobel, S. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, 18(1):7–35.

Bustamante, E. E., Rocha, A. M., Quincozes, S. E., Kazienko, J. F., and Quincozes, V. E. (2025). Caracterização de phishing com grandes modelos de linguagem (llms): Uma avaliação comparativa entre gemini, deepseek e chatgpt. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 204–215. SBC.

Chataut, R., Gyawali, P. K., and Usman, Y. (2024). Can ai keep you safe? a study of large language models for phishing detection. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 548–554. IEEE.

[1] `https://www.kaggle.com/datasets/subhajournal/phishingemails/`

Fette, I., Sadeh, N., and Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, pages 649–656. ACM.

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., and Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146.

Jakobsson, M. and Myers, S., editors (2006). *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. Wiley-Interscience.

Koide, T., Fukushi, N., Nakano, H., and Chiba, D. (2024a). Chatspamdetector: Leveraging large language models for effective phishing email detection. CoRR abs/2402.18093.

Koide, T., Nakano, H., and Chiba, D. (2024b). ChatPhishDetector: Detecting phishing sites using large language models. *IEEE Access*, 12:154381–154400.

Li, Y., Huang, C., Deng, S., Lock, M. L., Cao, T., Oo, N., Lim, H. W., and Hooi, B. (2024). KnowPhish: Large language models meet multimodal knowledge graphs for enhancing Reference-Based phishing detection. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 793–810, Philadelphia, PA. USENIX Association.

Lim, B., Huerta, R., Sotelo, A., Quintela, A., and Kumar, P. (2025). Explicate: Enhancing phishing detection through explainable ai and llm-powered interpretability. arXiv preprint arXiv:2503.20796.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Mahendru, S. and Pandit, T. (2024). Securenet: A comparative study of deberta and large language models for phishing detection. In *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)*, pages 160–169. IEEE.

Nahmias, D., Engelberg, G., Klein, D., and Shabtai, A. (2024). Prompted contextual vectors for spear-phishing detection. CoRR abs/2402.08309.

Nair, R., Abbasi, F. H., and Pervez, S. (2025). Phishemailllm: A meta model approach to detect phishing emails by leveraging llms and machine learning models. In *Proceedings of the 2025 Australasian Computer Science Week (ACSW 2025)*, pages 19–29. ACM.

Quincozes, C. B., Molinos, D., Araújo, R. D., and Quincozes, S. E. (2025). Indicadores semânticos na engenharia de prompts: Uma abordagem explicável para detecçao de fake news. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 1113–1120. SBC.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Sharevski, F., Devine, A., Pieroni, E., and Jachim, P. (2022). Phishing with malicious qr codes. In *Proceedings of the 2022 European Symposium on Usable Security*, pages 160–171. ACM.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., and Downs, J. (2010). Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373–382. ACM.

Subhajournal (2023). Phishing emails dataset. Kaggle dataset. Acesso em 11 maio 2026.

Uddin, M. A. and Sarker, I. H. (2024). An explainable transformer-based model for phishing email detection: A large language model approach. CoRR abs/2402.13871.