Caracterização de Phishing com Grandes Modelos de Linguagem (LLMs): Uma Avaliação Comparativa entre Gemini, DeepSeek e ChatGPT

Evelyn E. B. Bustamante[1], Adriano M. Rocha[2], Silvio E. Quincozes[2,3], Juliano F. Kazienko[4] e Vagner E. Quincozes[5]

[1]FACOM – Universidade Federal de Uberlândia (UFU) – Uberlândia, Brasil
[2]FACOM – Universidade Federal de Uberlândia (UFU) – Monte Carmelo, Brasil
[3]Universidade Federal do Pampa (UNIPAMPA) – Alegrete, Brasil
[4]CTISM – Universidade Federal de Santa Maria (UFSM) – Santa Maria, Brasil
[5]IC – Universidade Federal Fluminense (UFF) – Niterói, Brasil

{evelyn.bustamante, adriano.rocha}@ufu.br
silvioquincozes@unipampa.edu.br, kazienko@redes.ufsm.br,
vequincozes@id.uff.br

Abstract. *The rise in phishing attacks demands robust detection and characterization strategies. Large Language Models (LLMs) show promise in this domain, but their effectiveness—particularly that of newer models—remains unexplored. In this work, a novel phishing characterization method based on LLMs is proposed. Based on 1,009 analyzed emails, key phishing features were extracted by Gemini, DeepSeek, and ChatGPT using standardized prompts to ensure consistency in testing. Our results indicate that DeepSeek stands out in robustness and achieved the best overall performance, with an F1-Score of 92.38%.*

Resumo. *O aumento dos ataques de phishing demanda estratégias robustas de detecção e caracterização. Grandes Modelos de Linguagem (LLMs) mostram potencial nesse domínio, mas sua efetividade—especialmente a dos modelos mais recentes—ainda é pouco explorada. Neste trabalho, é proposto um novo método de caracterização de phishing baseado em LLMs. Com base em 1.009 e-mails analisados, foram extraídas características-chave pelos modelos Gemini, DeepSeek e ChatGP, utilizando prompts padronizados para garantir consistência nos testes. Nossos resultados indicam que o DeepSeek se destaca em robustez e apresentou o melhor desempenho geral, com um F1-Score de 92,38%.*

1. Introdução

A detecção de e-mails de *phishing* representa um desafio recorrente no campo da segurança digital, impulsionado pela sofisticação crescente das mensagens maliciosas e pela dificuldade de detectá-las [Ciso Advisor 2023]. Esses ataques utilizam técnicas de engenharia social, nas quais criminosos digitais enviam comunicações fraudulentas disfarçadas de fontes confiáveis para induzir usuários a revelar informações sensíveis ou comprometer sua segurança [IBM 2024]. A engenharia social, nesse contexto, se apoia mais em estratégias de manipulação psicológica do que em falhas técnicas, explorando a confiança do usuário para obter acesso não autorizado a sistemas e dados [Gov 2021].

O Brasil ocupa posição de destaque negativo nesse cenário: segundo a [CNN Brasil 2024], o país registra em média 1.379 ataques cibernéticos por minuto, sendo o *phishing* o tipo mais recorrente, com um custo médio de R$ 7,75 milhões por violação. Globalmente, o volume desses ataques também se mantém elevado: apenas no último trimestre de 2024, o *Anti-Phishing Working Group* contabilizou 989.123 incidentes [Associação APWG 2025]. Esses dados reforçam a urgência de soluções eficazes, escaláveis e automatizadas para mitigar o problema, especialmente diante da crescente complexidade das mensagens fraudulentas e das limitações dos filtros tradicionais.

Para combater essas ameaças, novas soluções são necessárias. Nesse contexto, os Grandes Modelos de Linguagem—do inglês, *Large Language Models* (LLMs)—têm se mostrado promissores ao reconhecer padrões linguísticos associados a tentativas de fraude. Os LLMs são fortemente baseados em técnicas de Processamento de Linguagem Natural (PLN), permitindo a extração de significados semânticos e contextuais de textos, fator crucial na detecção de mensagens de *phishing* [Oracle 2021]. Modelos como o ChatGPT, da OpenAI [OpenAI 2025], o Gemini, desenvolvido pelo Google DeepMind [Google 2023], e o DeepSeek, proposto por uma startup chinesa focada em aplicações técnicas e empresariais [BBC 2025], têm ampliado as fronteiras da IA em diferentes domínios, incluindo a cibersegurança. Cada um desses modelos apresenta características distintas quanto à arquitetura, volume de treinamento e capacidades de generalização, o que motiva investigações sobre sua eficácia em tarefas específicas, como a caracterização de e-mails maliciosos.

Estudos recentes vêm explorando o uso de LLMs na detecção de *phishing*, com diferentes abordagens e resultados. Por exemplo, Zhang et al. 2024 demonstraram bons índices de precisão na análise textual com o GPT-3.5-turbo, mas evidenciaram limitações na detecção de imagens e falhas após múltiplas iterações. Outros trabalhos como [Sayyafzadeh et al. 2024, Al Daoud et al. 2024, Chataut et al. 2024, Heiding et al. 2024, De Rosa et al. 2024] reforçam o potencial dos LLMs, mas apontam problemas como dependência de grandes volumes de dados, dificuldades de generalização e custo computacional elevado. Portanto, para além do emprego das LLMs na detecção de *phishing*, faz-se necessária uma abordagem estruturada que permita a caracterização de padrões linguísticos e de indícios comuns que funcionem como alertas para esse tipo de ataque [Ahmed 2024].

Este trabalho propõe uma nova estratégia de categorização de e-mails *phishing* baseada em LLMs, a qual concentra-se na busca por explicabilidade das decisões, mapeando como cada LLM interpreta e identifica padrões linguísticos associados a e-mails maliciosos. Para tanto, o método proposto se baseia na análise das características-chave (*red flags*) extraídas pelos modelos. Ademais, é realizada uma avaliação comparativa entre os modelos Gemini 2.0, ChatGPT 4 e DeepSeek 3.0. Como parte da contribuição metodológica, foi derivado um *dataset* contendo amostras de e-mails legítimos e maliciosos, as *red flags* extraídas por cada modelo e os resultados quantitativos das avaliações da intensidade dessas *red flags* em cada e-mail.

A organização deste trabalho está estruturada como descrito em seguida. Na Seção 2, é apresentada uma revisão da literatura acerca da temática estudada. A Seção 3 detalha a metodologia adotada na coleta, organização e análise dos dados. Na seção 4, é descrito o *dataset* gerado a partir das características de *phishing* apontadas pelas LLMs. Em seguida, a Seção 5 discute os resultados obtidos e suas implicações. Por fim, a Seção 6

apresenta as conclusões do estudo e propõe direções para trabalhos futuros.

2. Trabalhos Relacionados

Diversos estudos recentes têm explorado abordagens inovadoras para a detecção de e-mails de *phishing* com o uso de modelos de linguagem natural. A seguir, os trabalhos são discutidos considerando similaridade metodológica e abordagem tecnológica. Ademais, são destacadas suas contribuições, limitações e lacunas relevantes para o presente estudo.

2.1. Uso de LLMs na detecção de *phishing*

Estudos como [De Rosa et al. 2024, Zhang et al. 2024, Beydemir et al. 2024, Chataut et al. 2024, Sayyafzadeh et al. 2024] concentraram suas investigações em modelos da família GPT para interpretar e classificar e-mails maliciosos. Esses trabalhos variam desde a simples aplicação de modelos pré-treinados até a combinação com estratégias complementares, como análise de sentimentos e módulos de decisão adaptativa. É importante destacar que os trabalhos discutidos nesta seção focam na avaliação de uma LLM em particular, sem, no entanto, investigar a eficiência de classificação e consequente comparação de um rol de LLMs diferentes.

De Rosa et al. 2024 propôs uma solução que integra análise estática (cabeçalhos, corpo, links e anexos) com análise dinâmica via GPT-3.5. Os testes com 2.000 e-mails revelaram precisão de 75,75% e sensibilidade de 98,4%, mas baixa especificidade (53,1%). O modelo teve destaque na análise de anexos, embora apresentasse limitações como tempo de processamento elevado, dependência de APIs com falhas e cobertura parcial dos componentes de e-mails. Por sua vez, Zhang et al. 2024 investigou o uso de múltiplas versões do ChatGPT (3.5 e 4.0) em uma ferramenta automatizada de detecção, combinando listas restritivas e aprendizado de máquina com 11.430 URLs e 87 atributos. O desempenho foi promissor na análise textual e de URLs, mas limitado na detecção de imagens maliciosas e sujeito a inconsistências após múltiplas interações, além de enfrentar dificuldades na extração de dados e problemas de estabilidade de rede.

Beydemir et al. 2024 desenvolveu um sistema que seleciona dinamicamente o melhor modelo GPT para cada instância de e-mail com base em histórico de desempenho. Usando *datasets* do Enron [Cohen 2015] e conjuntos sintéticos, o modelo *fine-tuning* obteve até 96% de revocação. Apesar do alto desempenho, o sistema apresenta limitações como dependência de treinamento específico, suscetibilidade a falsos positivos em contextos técnicos e custos operacionais elevados devido ao uso do GPT-4.

De forma complementar, Chataut et al. 2024 avaliou os modelos GPT-3.5, GPT-4 e um modelo personalizado (CyberGPT), usando 828 e-mails. O GPT-4 e o CyberGPT superaram 97% de acurácia, enquanto o GPT-3.5 ficou em 80,68%. O CyberGPT se destacou por identificar padrões de linguagem urgentes e remetentes suspeitos, mas os autores alertam sobre sua vulnerabilidade a falsos negativos, necessidade de atualização constante e elevado custo computacional.

Sayyafzadeh et al. 2024 integraram LLMs com análise de sentimentos usando a ferramenta VADER. O sistema atingiu 92% de acurácia ao explorar não apenas o conteúdo textual, mas também emoções como urgência e medo, demonstrando potencial em contextos corporativos. Ainda assim, desafios como dependência de modelos caros e adaptação a idiomas menos comuns permanecem.

Ademais, Jiang 2024 examina o uso de LLMs (GPT-3.5/4) para detectar golpes cibernéticos via análise de padrões linguísticos, utilizando metodologia de coleta diversificada, pré-processamento e *fine-tuning*, com resultados preliminares mostrando eficiência na identificação de sinais como erros gramaticais e links suspeitos, embora apresente limitações significativas: análise restrita a um único caso de *phishing*, falta de diversidade de ataques, ausência de métricas quantitativas robustas e avaliação em ambientes reais, deixando questões sobre escalabilidade e adaptação a novas ameaças sem resposta adequada para validação prática em segurança cibernética.

2.2. Comparação entre LLMs e Abordagens Alternativas

Trabalhos como [Al Daoud et al. 2024] e [Heiding et al. 2024] analisam o uso de LLMs, como GPT-4o, Claude, Gemini, Bard e LLaMA-2, na detecção de *phishing* em e-mails e redes sociais. Foram comparadas abordagens como *zero-shot learning*, extração de características, *fine-tuning* e *ensemble*, com destaque para os LLMs, que alcançaram até 99,21% de acurácia. Claude-1 obteve 100% de acerto em e-mails sofisticados, enquanto modelos como Bard e LLaMA-2 apresentaram desempenho inferior. Também foi avaliado o uso de LLMs para criar ataques, sendo os métodos manuais ainda mais eficazes. Apesar do bom desempenho, os estudos apontam limitações importantes, como viés linguístico, alta dependência de *prompts*, variabilidade nos resultados, falta de explicabilidade e vulnerabilidade a novos ataques — fatores que limitam sua aplicação prática e substituição de especialistas humanos.

Outros trabalhos propõem métodos distintos dos modelos de linguagem de última geração. Em [Xu et al. 2014], uma extensão de navegador (Gemini) foi implementada para bloquear sites fraudulentos ao detectar o uso de credenciais em domínios não reconhecidos. Com 0% de falsos negativos e menos de 1% de falsos positivos, a solução mostrou-se leve e eficaz. Contudo, é limitada a páginas padronizadas e não protege nomes de usuário e enfrenta dificuldades com tecnologias como Flash.

Em [Ramprasath et al. 2023], redes neurais recorrentes com LSTM foram aplicadas para detectar padrões temporais em e-mails. A solução superou métodos como SVM e k-NN, destacando-se por sua adaptabilidade. No entanto, apresenta alto custo computacional, sensibilidade a ruídos e pouca interpretabilidade.

2.3. Síntese comparativa

A Tabela 1 apresenta uma comparação entre os trabalhos relacionados e este estudo. São analisados o escopo de aplicação (como *phishing* em e-mails ou redes sociais), os modelos de linguagem investigados (incluindo GPT, Gemini, Claude, DeepSeek e LLaMA), a presença de caracterização de *phishing* e a geração de novos conjuntos de dados.

Tabela 1. Comparação entre trabalhos relacionados e o presente estudo.

Referência	Escopo	Modelos Avaliados					Caracterização de Phishing	Dataset
		GPT	Gemini	Claude	DeepSeek	LLaMA		
De Rosa et al. 2024	Phishing (e-mail)	3.5	–	–	–	–	–	–
Zhang et al. 2024	Phishing (e-mail)	3.5 / 4	–	–	–	–	–	–
Beydemir et al. 2024	Phishing (e-mail)	4 + FT	–	–	–	–	–	–
Al Daoud et al. 2024	Phishing (e-mail, rede social)	4	1.5	3	–	–	–	–
Sayyafzadeh et al. 2024	Phishing (e-mail)	4	–	–	–	–	–	–
Chataut et al. 2024	Phishing (e-mail)	3.5, 4, CyberGPT	–	–	–	–	–	–
Heiding et al. 2024	Phishing (Gera e detecta)	4	✓	1	–	2	–	–
Jiang 2024	Phishing (e-mail)	3.5 / 4	–	–	–	–	✓	–
Este trabalho	Phishing (e-mail)	4	2	–	3	–	✓	✓

A coluna "Modelos Avaliados" indica quais variantes de LLMs foram exploradas em cada estudo, com a sigla "–" sinalizando ausência de análise sobre determinado modelo e o símbolo "✓" representando sua inclusão, mesmo quando a versão exata não é especificada, como no caso de Gemini em Heiding et al. 2024. Já a coluna "Caracterização de Phishing" mostra se houve extração de características-chave (*red flags*) nos e-mails, sendo essa uma das principais contribuições metodológicas do presente trabalho. A coluna "Dataset" indica se o estudo resultou na criação de um novo conjunto de dados baseado nas respostas dos modelos.

Diferentemente das abordagens anteriores, este estudo não apenas avalia múltiplos LLMs em um mesmo cenário controlado, como também propõe um método de caracterização automatizada de e-mails de *phishing* e gera um conjunto de dados enriquecido a partir dessas análises.

3. Materiais e Métodos

Este estudo foi conduzido em etapas distintas, abrangendo a definição do *dataset* de e-mails adotado, a padronização e organização dos dados e a extração de características que identificam *phishing*. O novo *dataset*, o qual é resultante da metodologia aplicada nesta pesquisa, é apresentado na Seção 4.

3.1. *Dataset* Adotado

O conjunto de dados utilizado foi obtido no trabalho de [Chakraborty 2023], escolhido pela clareza na estrutura das informações e pela elevada avaliação (9.41) na plataforma Kaggle[1]. Os e-mails presentes no conjunto estão escritos em inglês. No total, 97% apresentam corpo textual e 3% contêm dados ausentes. Em relação à classificação, 61% são

[1]Plataforma Kaggle, disponível em: `https://www.kaggle.com/datasets/subhajournal/phishingemails/data?select=Phishing_Email.csv`

e-mails legítimos (*Safe Email*) e 39% são e-mails maliciosos (*Phishing Email*). O conjunto de dados é composto por duas colunas principais: ***E-mail Text***, denotando o corpo textual do e-mail e ***E-mail Type***, denotando o tipo do e-mail, ou seja, *Phishing* ou Seguro. A partir deste *dataset*, foram utilizadas 1.009 amostras, número definido com base na média observada em trabalhos relacionados, o que permite uma comparação adequada dos resultados. Importante destacar que a amostra reproduz a mesma distribuição proporcional do *dataset* total, mantendo a representatividade dos e-mails legítimos e maliciosos. Desse total, 602 correspondem a e-mails legítimos (*Safe E-mail*) e 407 a e-mails maliciosos (*Phishing E-mail*).

3.2. Extração de Características Chave de *Phishing*

Foram utilizados três sistemas distintos de inteligência artificial generativa: Gemini, DeepSeek e ChatGPT, utilizando o mesmo *prompt* para todos os modelos a fim de garantir consistência na comparação dos resultados.

Os primeiros 10 e-mails foram apresentados a cada IA com os comandos: Os primeiros 10 e-mails foram submetidos a cada IA junto aos respectivos comandos. Desses, 6 são legítimos e 4 são e-mails de *phishing*.

1. "Analise o conteúdo abaixo e identifique os elementos que ajudam a identificar um e-mail de phishing. Ao final, liste todos esses elementos de forma breve em forma de itens com uma frase simples explicando".
2. "Agora crie uma planilha onde cada um desses elementos é uma das colunas e marque para cada um dos itens o valor **SIM** ou **NÃO**".

Cada LLM produziu um conjunto distinto de colunas representando características potenciais de *phishing* nos e-mails analisados. Para ampliar o escopo, expandimos a análise para uma amostra adicional de 100 e-mails composta por 62 e-mails legítimos e 38 e-mails de *phishing*, solicitando aos modelos que identificassem características ainda não catalogadas. Apesar da análise para uma amostra adicional de 100 e-mails, durante a identificação de características de *phishing*, observou-se que as LLMs deixaram de produzir novos atributos relevantes após um determinado ponto, apontando para a ausência de novas características relevantes. Em virtude da redundância observada nas características geradas, a análise foi limitada às identificadas nos primeiros 100 e-mails, preservando um conjunto conciso e representativo. Características repetidas foram removidas, resultando em conjuntos otimizados e distintos de características *(features)* para cada LLM. Na Tabela 2, são apresentadas as características indicadas pelas LLMs abordadas neste trabalho.

Além disso, pedimos às LLMs que criassem uma coluna chamada "Probabilidade de *Phishing*", utilizando o *prompt*: "Adicione à tabela: Probabilidade de Phishing e seu resultado", com três níveis de classificação baseados nas características de *phishing* identificadas. A seguir, são apresentados os três níveis: **BAIXA:** Indica que o e-mail provavelmente não é *phishing*; **MÉDIA:** Indica possibilidade de *phishing*; e **ALTA:** Indica alta probabilidade de ser *phishing*.

A probabilidade foi determinada com base no número de características de *phishing* presentes em cada e-mail. Se um e-mail apresentar 5 ou mais elementos marcados como "SIM", a probabilidade é classificada como Alta. E-mails com entre 2 e 4

Característica	Descrição	Gemini	ChatGPT	DeepSeek
Remetente suspeito	Remetente é desconhecido, nome genérico ou parece forjado.	X	X	X
Senso de urgência ou medo	Golpistas criam um senso de urgência para que o usuário aja rapidamente, sem pensar.	X	X	X
Solicitação de informações sensíveis	Remetente induz o destinatário a fornecer dados sensíveis, como senhas ou informações bancárias.	X	X	X
***Links* suspeitos**	*Links* em e-mails que aparentam direcionar para sites legítimos, mas que levam a páginas falsas ou maliciosas.	X	X	X
Erros gramaticais e ortográficos	Textos com muitos erros de gramática e ortografia.	X	X	X
E-mail não solicitado	Mensagem recebida sem que o destinatário tenha se inscrito para recebê-la.	X	X	
Saudação genérica	Saudação não dirigida especificamente a pessoa.	X	X	
Anexos suspeitos	Arquivos enviados por e-mail que podem conter *malware*.	X	X	
Formatação estranha	Erros visuais no e-mail, como espaçamento irregular, fontes diferentes ou imagens de baixa qualidade.	X	X	
Oferta boa demais	Mensagens que prometem benefícios excessivos, como dinheiro fácil.	X		X
Domínio suspeito	E-mail cujo domínio não corresponde exatamente ao domínio oficial da organização que supostamente enviou a mensagem.	X		X
Conteúdo impróprio	E-mail com conteúdo sexualmente explícito.	X		
Remetente falsificado	Golpistas forjam o endereço do remetente para parecer que o e-mail vem de uma fonte confiável.	X		
Histórias elaboradas	Narrativas complexas ou emocionais que tentam convencer a vítima a agir sem cautela, como fornecer dados pessoais.	X		
Personalização excessiva	Golpistas utilizam informações pessoais do destinatário, como nome ou cargo na empresa, para criar um e-mail que pareça legítimo, aumentando a confiança.	X		
Falta de informação de contato	E-mail que não apresenta dados básicos como telefone ou endereço da empresa.	X		
Conteúdo emocionalmente apelativo	Mensagens que despertam emoções fortes, como medo, urgência, pena ou empatia, com o objetivo de manipular o destinatário.	X		
Informações de contato inconsistentes	Inconsistências entre os dados fornecidos no e-mail (como telefone ou endereço) e as dados oficiais presentes no site da empresa.	X		
Anexos com nomes estranhos	Anexos em e-mails que apresentam nomes genéricos ou fora do contexto.	X		
Endereço de resposta diferente	O endereço para o qual as respostas do e-mail são enviadas ("Reply-To") é diferente do endereço original do remetente.		X	
Falta de personalização	Mensagens genéricas que não usam o nome ou informações específicas do usuário alvo do ataque.			X
Probabilidade de *Phishing*	Estabelecida em função das características anteriores.	X	X	X

Tabela 2. Características de *phishing* e-mails indicadas pelas LLMs estudadas.

elementos são classificados com probabilidade Média, enquanto e-mails com menos de 2 elementos são classificados com probabilidade Baixa. Por fim, qualquer e-mail classificado com probabilidade Média ou Alta foi considerado como *phishing*.

Como observado na Tabela 2, todos os LLMs identificaram as características: *remetente suspeito*, *senso de urgência ou medo*, *solicitação de informações sensíveis*, *links suspeitos* e *erros gramaticais e ortográficos*. Isso indica que essas características são frequentemente utilizadas em e-mails de *phishing*. Por sua vez, as características *e-mail não solicitado*, *saudação genérica*, *anexos suspeitos* e *formatação estranha* foram reconhecidas pelo Gemini e pelo ChatGPT. Já as características *oferta boa demais* e *domínio suspeito* foram identificadas pelo Gemini e pelo DeepSeek.

Algumas características foram reconhecidas exclusivamente pelo Gemini, tais como: *conteúdo impróprio*, *remetente falsificado*, *histórias elaboradas*, *personalização excessiva*, *falta de informação de contato*, *conteúdo emocionalmente apelativo*, *informações de contato inconsistentes* e *anexos com nomes estranhos*. Embora o Gemini tenha identificado um maior número de características, algumas delas são redundantes, como *remetente suspeito* e *remetente falsificado*.

A característica *endereço de resposta diferente* foi identificada apenas pelo ChatGPT, enquanto a característica *falta de personalização* foi reconhecida exclusivamente pelo DeepSeek.

4. *Dataset* Gerado

Como contribuição metodológica adicional deste estudo, desenvolveu-se um *dataset* derivado que amplia substancialmente o potencial analítico da investigação. Este conjunto de dados vai além dos e-mails originais testados (*dataset* adotado), incluindo as características de *phishing* que estão na Tabela 2. Tais características foram selecionadas com base na identificação feita pelo ChatGPT, considerando que a maioria também foi apontada pelo Gemini e/ou DeepSeek. Além disso, para cada e-mail analisado, registrou-se tanto os resultados da classificação quanto as métricas de desempenho, permitindo entender melhor o comportamento dos modelos.

Com este material produzido, é possível comparar os diferentes métodos de detecção de maneira clara, identificando onde cada abordagem se destaca e onde apresenta dificuldades. O *dataset* completo produzido neste trabalho está disponibilizado no GitHub[2], o que oportuniza a reprodutibilidade dos resultados obtidos nesta pesquisa.

5. Resultados e Discussão

Nesta seção, apresentamos e analisamos os resultados dos modelos avaliados (Gemini, DeepSeek e ChatGPT) com base nas métricas de desempenho: Precisão, *Recall*, Acurácia e F1-Score, conforme ilustrado na Figura 1. Os resultados evidenciam diferenças significativas no comportamento dos modelos, especialmente no contexto da detecção de *phishing*, uma tarefa que demanda atenção tanto à identificação precisa de ameaças quanto à minimização de alertas indevidos.

[2]Download do *Dataset* Gerado: `https://github.com/evelyn-bustamante/Phishing_Analise_Resultados`

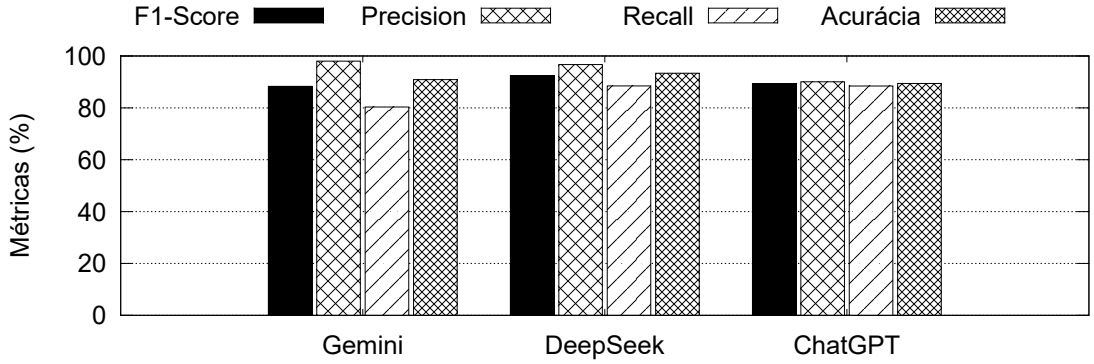

Figura 1. Resultados consolidados das métricas de desempenho dos modelos Gemini, DeepSeek e ChatGPT: Acurácia, Precisão, *Recall* e F1-Score.

O DeepSeek obteve o melhor desempenho geral, com F1-Score de 92,38%, precisão de 96,68%, *recall* de 88,45% e acurácia de 93,36%. Esses resultados indicam um classificador robusto, capaz de detectar mensagens suspeitas com eficácia sem comprometer a confiabilidade das decisões. Em cenários de *phishing*, onde tanto a falha em detectar ameaças reais (falsos negativos) quanto a sinalização indevida de mensagens legítimas (falsos positivos) podem gerar impactos significativos, esse equilíbrio torna o DeepSeek uma solução adequada para ambientes que exigem proteção ampla e com maior consistência que os demais modelos.

O Gemini destacou-se pela alta precisão, atingindo 98,01%, o que evidencia sua capacidade de reduzir falsos alarmes, essencial em contextos onde respostas automatizadas a detecções de *phishing* podem acarretar bloqueios ou prejuízos operacionais se acionadas indevidamente. No entanto, sua *recall* de 80,34% sugere uma menor sensibilidade à detecção completa de tentativas de *phishing*, o que implica que algumas ameaças podem passar despercebidas. Dessa forma, o Gemini é mais indicado em ambientes onde a confiança nas classificações positivas deve ser maximizada, mesmo que se aceite um nível reduzido de cobertura na detecção.

O ChatGPT apresentou métricas intermediárias, com F1-Score de 89,23%, precisão de 90,05%, *recall* de 88,42% e acurácia de 89,40%. Sua performance equilibrada entre precisão e *recall* o posiciona como uma alternativa versátil, especialmente em cenários onde é necessário manter uma boa capacidade de resposta frente a mensagens maliciosas, sem comprometer excessivamente a experiência do usuário com alertas falsos. Embora não alcance os extremos de desempenho observados nos outros modelos, sua consistência o torna uma escolha sólida para aplicações de detecção de *phishing* que não demandem rigidez extrema em uma métrica específica.

Portanto, a comparação entre os modelos demonstra que a escolha da abordagem mais adequada está diretamente relacionada ao perfil de risco e aos objetivos operacionais da aplicação. Em sistemas onde o impacto de falsos positivos deve ser minimizado, o Gemini é a opção preferencial. Já em contextos que demandam a máxima detecção de ameaças, mesmo com um custo moderado em confiabilidade, o DeepSeek sobressai. O ChatGPT, por sua vez, atende bem a cenários onde se busca um equilíbrio prático entre sensibilidade e precisão, sendo uma solução intermediária confiável. Dessa forma, a principal constatação é que ao implantar-se uma solução de detecção de *phishing* é funda-

mental considerar quais são os requisitos prioritários da organização ou do usuário final do sistema de e-mails.

6. Conclusão

Este estudo realizou uma avaliação comparativa do desempenho dos modelos de LLMs Gemini, DeepSeek e ChatGPT na tarefa de detecção de e-mails de *phishing*. A análise dos dados demonstrou que, embora todos os modelos tenham apresentado resultados satisfatórios, houve variação nos valores das métricas de desempenho obtidas.

Os resultados indicaram que o DeepSeek alcançou os melhores índices, destacando-se em F1-Score (92,38%), Acurácia (93,36%) e *Recall* (88,45%), superando os demais modelos. O Gemini, apesar de registrar a maior Precisão (98,01%), apresentou limitações na *Recall* (80,34%), enquanto o GPT não se destacou em nenhuma métrica específica, obtendo Acurácia de 89,40%, Precisão de 90,05%, *Recall* de 88,42% e F1-Score de 89,23%, porém apresentou F1-Score e *Recall* superiores aos do Gemini.

Conforme evidenciado pelos resultados, o Gemini se destaca como um modelo altamente confiável para predições positivas, sendo ideal quando a prioridade é minimizar falsos alarmes. O ChatGPT apresenta um desempenho equilibrado, embora não tenha se destacado em nenhuma métrica específica. Já o DeepSeek pode ser a melhor escolha quando o objetivo é maximizar a eficiência global, especialmente em cenários sensíveis que exigem equilíbrio entre Precisão e *Recall*.

Por fim, além das contribuições metodológicas e experimentais, este trabalho também disponibiliza um *dataset* enriquecido, contendo *red flags* identificadas pelas LLMs. Esse recurso está disponível publicamente e pode servir de base para futuras pesquisas na interseção entre segurança cibernética e inteligência artificial generativa.

Como possibilidades para trabalhos futuros, propõem-se diversas linhas de aprofundamento. Primeiramente, sugere-se a ampliação dos experimentos, com a realização de novos testes para aumentar a robustez estatística dos resultados obtidos. Outra direção promissora é a expansão multilíngue, por meio da extensão da análise para e-mails de *phishing* em diferentes idiomas, permitindo avaliar a consistência dos padrões de detecção frente a distintas estruturas linguísticas. Além disso, recomenda-se uma investigação mais aprofundada dos fatores que influenciam diretamente o *recall*, como o tamanho e a complexidade dos e-mails, bem como a presença de padrões linguísticos específicos associados a *red flags* de *phishing*. Também se considera relevante a realização de uma análise comparativa expandida, incluindo no estudo outras LLMs emergentes, de modo a oferecer uma visão mais abrangente sobre o desempenho da IA generativa na detecção de *phishing*. Por fim, propõe-se o refinamento das técnicas de *prompt engineering*, explorando estratégias mais sofisticadas de construção de *prompts*, com o objetivo de potencializar a eficácia dos modelos nessa tarefa.

Referências

Ahmed (2024). Large Language Models (LLMs) in Cybersecurity: A Paradigm Shift in Threat Intelligence. Disponível em: `https://mawgoud.medium.com/large-language-models-llms-in-cybersecurity-a-paradigm-shift-in-threat-intelligence-5e0f7653dc11`. Acessado em 01 de maio de 2025.

Al Daoud, E., Al Daoud, L., Asassfeh, M., Al-Shaikh, A., Al-Sherideh, A. S., and Afaneh, S. (2024). Enhancing cybersecurity with transformers: Preventing phishing emails and social media scams. In *2024 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 31–36.

Associação APWG (2025). Relatórios de tendências de atividades de phishing. Disponível em: https://apwg.org/trendsreports/. Acessado em 09 de abril de 2025.

BBC (2025). DeepSeek: O aplicativo de IA chinês que está dando o que falar no mundo. Disponível em: https://www.bbc.com/news/articles/c5yv5976z9po. Acessado em 03 de abril de 2025.

Beydemir, A. B., Sezgin, U., Doğan, U., Aşıklar, B. E., Yerlikaya, F. A., and Bahtiyar, Ş. (2024). A dynamically selected gpt model for phishing detection. In *2024 14th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 481–484.

Chakraborty, S. (2023). Phishing Email Detection. Disponível em: https://www.kaggle.com/dsv/6090437. Acessado em 03 de abril de 2025.

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 0548–0554.

Ciso Advisor (2023). E-mail de phishing gerado por IA é quase impossível de detectar. Disponível em: https://www.cisoadvisor.com.br/e-mail-de-phishing-gerado-por-ia-e-quase-impossivel-de-detectar/. Acessado em 01 de maio de 2025.

CNN Brasil (2024). Brasil é vice-campeão em ataques cibernéticos, com 1.379 golpes por minuto, aponta estudo. Disponível em: https://www.cnnbrasil.com.br/economia/negocios/brasil-e-vice-campeao-em-ataques-ciberneticos-com-1-379-golpes-por-minuto-aponta-estudo/. Acessado em 28 de março de 2025.

Cohen, W. W. (2015). Enron Email Dataset. Disponível em: https://www.cs.cmu.edu/~enron/. Universidade Carnegie Mellon. Acessado em 25 de maio de 2025.

De Rosa, S., Gringoli, F., and Bellicini, G. (2024). Hey chatgpt, is this message phishing? In *2024 22nd Mediterranean Communication and Computer Networking Conference (MedComNet)*, pages 1–10.

Google (2023). Apresentando o Gemini: nosso maior e mais hábil modelo de IA. Disponível em: https://blog.google/intl/pt-br/novidades/tecnologia/apresentando-o-gemini-nosso-maior-e-mais-habil-modelo-de-ia. Acessado em 03 de abril de 2025.

Gov, P. (2021). Engenharia social. Guia para Proteção de Conhecimentos Sensíveis. Disponível em: https://www.gov.br/abin/pt-br/institucional/acoes-e-programas/PNPC/boaspraticas/EngenhariaSocial_27062022.pdf. Acessado em 01 de abril de 2025.

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.

IBM (2024). O que é phishing? Disponível em: https://www.ibm.com/br-pt/topics/phishing. Acessado em 01 de abril de 2025.

Jiang, L. (2024). Detecting scams using large language models. *arXiv preprint arXiv:2402.03147*.

OpenAI (2025). O que é o ChatGPT? Disponível em: https://chatgpt.com.br/faq/#o-que-e-chatgpt. Acessado em 03 de abril de 2025.

Oracle (2021). O que é Processamento de Linguagem Natural (PLN)? Disponível em: https://www.oracle.com/br/artificial-intelligence/what-is-natural-language-processing/. Acessado em 03 de abril de 2025.

Ramprasath, J., Priyanka, S., Manudev, R., and Gokul, M. (2023). Identification and mitigation of phishing email attacks using deep learning. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 466–470.

Sayyafzadeh, S., Weatherspoon, M., Yan, J., and Chi, H. (2024). Securing against deception: Exploring phishing emails through chatgpt and sentiment analysis. In *2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 159–165.

Xu, Z., Wang, H., and Jajodia, S. (2014). Gemini: An emergency line of defense against phishing attacks. In *2014 IEEE 33rd International Symposium on Reliable Distributed Systems*, pages 11–20.

Zhang, D., Jain, K., and Singh, P. (2024). Guarding against chatgpt threats: Identifying and addressing vulnerabilities. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 612–615.