

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Evelyn Estefania Bravo Bustamante

**Uma Metodologia Comparativa para Avaliação
de LLMs na Detecção de Phishing: Um Estudo
Envolvendo DeepSeek, Gemini e ChatGPT**

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Evelyn Estefania Bravo Bustamante

**Uma Metodologia Comparativa para Avaliação de LLMs
na Detecção de Phishing: Um Estudo Envolvendo
DeepSeek, Gemini e ChatGPT**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Orientador: Prof. Dr. Adriano Mendonça Rocha

Coorientador: Prof. Dr. Silvio Ereno Quincozes

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

Evelyn Estefania Bravo Bustamante

Uma Metodologia Comparativa para Avaliação de LLMs na Detecção de Phishing: Um Estudo Envolvendo DeepSeek, Gemini e ChatGPT

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 09 de maio de 2025:

Prof. Dr. Adriano Mendonça Rocha
Orientador

Prof. Dr. Silvio Ereno Quincozes
Coorientador

Prof. Dr. Diego Nunes Molinos
Convidado 1

Prof. Giullia Rodrigues de Menezes
Convidado 2

Uberlândia, Brasil
2025

Agradecimentos

Agradeço aos meus pais, por me ensinarem valores de perseverança e resiliência.

Aos meus amigos e companheiro que me apoiaram com palavras de bravura e conforto.

Às pessoas que participaram da minha caminhada dentro e fora da sala de aula.

Ao meus orientadores, pela paciência e ensinamentos.

E, por fim, à Faculdade de Computação da UFU, que me proporcionou uma bagagem de conhecimentos e experiências que me acompanharão por toda a vida.

*“Ninguém é tão grande que não possa aprender,
nem tão pequeno que não possa ensinar.”*

(Esopo)

Resumo

A detecção de *phishing* representa um desafio crescente no campo da segurança digital, devido à sofisticação das mensagens maliciosas e à dificuldade em identificar ataques de maneira automatizada e confiável. Embora Grandes Modelos de Linguagem (LLMs) tenham sido investigados como potenciais soluções, sua eficácia ainda precisa ser validada em cenários reais, especialmente para modelos mais recentes. Este estudo propõe uma avaliação comparativa entre os modelos Gemini, DeepSeek e ChatGPT, focando na classificação de e-mails legítimos e maliciosos. Com base em 1.009 amostras analisadas, o estudo caracterizou os e-mails com base nas *red flags* (características-chave) extraídas por cada modelo, e aplicou *prompts* padronizados para garantir consistência nos testes. As métricas de avaliação incluíram precisão, *recall*, acurácia e F1-Score, complementadas por uma análise qualitativa das respostas. Os resultados indicam que o DeepSeek se destaca em robustez, e obteve o melhor desempenho geral, com F1-Score de 92,38%.

Palavras-chave: Gemini, DeepSeek, ChatGPT, Detecção de E-mails de Phishing, Grandes Modelos de Linguagem

Lista de ilustrações

Figura 1 – Principais etapas do desenvolvimento.	23
Figura 2 – Resultados com o Gemini considerando o valor “nenhuma”, na coluna “Probabilidade de Phishing”	29
Figura 3 – Resultados com o Gemini após a alteração das amostras classificadas como “nenhuma” para “baixa”, na coluna “Probabilidade de Phishing”	30
Figura 4 – Resultados do DeepSeek	31
Figura 5 – Resultados do ChatGPT	32
Figura 6 – Resultados consolidados das métricas de desempenho dos modelos Ge- mini, DeepSeek e ChatGPT	32

Lista de tabelas

Tabela 1 – Comparação entre trabalhos relacionados e o presente estudo.	22
Tabela 2 – Características comuns de e-mails de <i>phishing</i> detectados pelo Gemini (Parte 1 de 2).	25
Tabela 3 – Características comuns de e-mails de <i>phishing</i> detectados pelo Gemini (Parte 2 de 2).	26
Tabela 4 – Características comuns de e-mails de <i>phishing</i> detectado o DeepSeek. .	26
Tabela 5 – Características comuns de e-mails de <i>phishing</i> detectado o ChatGPT. .	27
Tabela 6 – Matriz de Confusão do Gemini considerando o valor “nenhuma”, na coluna “Probabilidade de Phishing”.	30
Tabela 7 – Matriz de Confusão do Gemini após a alteração das amostras classi- ficadas como “nenhuma” para “baixa”, na coluna “Probabilidade de Phishing”.	30
Tabela 8 – Matriz de Confusão do DeepSeek	31
Tabela 9 – Matriz de Confusão do ChatGPT	32

Lista de abreviaturas e siglas

IA	Inteligência Artificial
LLM	<i>Large Language Models</i> (Grandes Modelos de Linguagem)
NLP	<i>Natural Language Processing</i> (Processamento de Linguagem Natural)
TXT	Arquivo de texto
CSV	<i>Comma-Separated Values</i>
API	<i>Application Programming Interface</i> (Interface de Programação de Aplicação)
LSTM	<i>Long Short-Term Memory</i>
RNN	Rede Neural Recorrente
SVM	<i>Support Vector Machine</i>
URL	<i>Uniform Resource Locator</i> (Localizador Uniforme de Recursos)

Sumário

1	INTRODUÇÃO	11
1.1	Objetivos	12
1.2	Justificativa	13
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	Conceitos Básicos	15
2.1.1	Phishing e Engenharia Social	15
2.1.2	Inteligência Artificial e Aprendizado de Máquina	15
2.1.3	Modelos de IA Generativa e LLMs	16
2.1.4	Processamento de Linguagem Natural (PLN)	16
2.1.5	Ferramentas baseadas em LLMs	16
2.2	Trabalhos Relacionados	17
3	DESENVOLVIMENTO	23
3.1	Dataset Adotado	23
3.2	Organização dos dados	24
3.3	Extração de Características Chave de <i>Phishing</i>	24
3.3.1	Ajustes Finais	24
3.3.2	Gemini	25
3.3.3	DeepSeek	26
3.3.4	ChatGPT	26
3.4	Avaliação e Comparação de Resultados	27
3.5	Dataset Gerado	28
4	RESULTADOS	29
4.1	Gemini	29
4.2	DeepSeek	31
4.3	ChatGPT	31
4.4	Comparações	32
5	CONCLUSÃO	34
	REFERÊNCIAS	35

APÊNDICES	39
APÊNDICE A – ACESSO AO DATASET	40

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 ([CISCO, 2024](#)). Nesse contexto, ferramentas baseadas em Inteligência Artificial, como os Grandes Modelos de Linguagem — do inglês, *Large Language Models* (LLMs) — têm se mostrado promissoras ao reconhecer padrões linguísticos associados a tentativas de fraude. No entanto, requer uma abordagem estruturada e colaborativa que integre soluções técnicas, estruturas e esforços contínuos de pesquisa ([MEDIUM, 2024](#)).

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.

Estudos recentes vêm explorando o uso de LLMs na detecção de *phishing*, com diferentes abordagens e resultados. [Zhang, Jain e Singh \(2024\)](#) 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. Em ([ROSA; GRINGOLI; BELLICINI, 2024](#)), os autores propuseram um sistema que combina análise estática (examinando cabeçalhos, corpo do texto, links e anexos) e análise dinâmica (utilizando o modelo para interpretação e classificação), com 75,75% de precisão, embora com alto tempo de resposta e baixa aplicabilidade prática. [Beydemir et al. \(2024\)](#) testaram seleção dinâmica de modelos, mas com fraca atuação em mensagens curtas. Outros trabalhos como ([SAYYAFZADEH et al., 2024](#); [DAOUD et al., 2024](#); [CHATAUT](#); [GYAWALI](#); [USMAN, 2024](#); [HEIDING 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. [Jiang \(2024\)](#) avaliou LLMs (GPT-3.5/GPT-4) na detecção de golpes cibernéticos via análise linguística, com resultados iniciais positivos, embora com limitações metodológicas e falta de validação prática. Assim, permanece a necessidade de investigações mais controladas e comparativas, especialmente com modelos mais recentes, como o DeepSeek, para o qual não foram encontrados estudos relacionados ao contexto

deste trabalho.

Este trabalho propõe uma avaliação comparativa entre as LLMs Gemini, ChatGPT e o modelo emergente DeepSeek. O principal objetivo é entender como cada modelo interpreta e identifica padrões linguísticos associados a e-mails maliciosos, por meio da análise das características-chave (*red flags*) extraídas pelos modelos. A avaliação é realizada com base em 1.009 amostras, utilizando *prompts* padronizados para garantir a consistência dos testes. Além disso, o desempenho dos modelos foi analisado por meio das métricas de precisão, *recall*, acurácia e F1-Score, complementadas por uma análise qualitativa do desempenho de cada modelo. Como parte da contribuição metodológica, foi desenvolvido um *dataset* derivado, contendo amostras de e-mails legítimos e maliciosos, as *red flags* extraídas por cada modelo e os resultados quantitativos das avaliações. Este estudo também observa o comportamento dos modelos frente a variações no tamanho dos textos dos e-mails, identificando inconsistências específicas em alguns deles.

A organização deste trabalho está estruturada como descrito em seguida. O Capítulo 2 apresenta os principais conceitos teóricos sobre *phishing* e LLMs. Na Seção 2.2, é apresentada uma revisão da literatura acerca da temática estudada. O Capítulo 3 detalha a metodologia adotada na coleta, organização e análise dos dados. Em seguida, o Capítulo 4 discute os resultados obtidos e suas implicações. Por fim, o Capítulo 5 apresenta as conclusões do estudo e propõe direções para trabalhos futuros.

1.1 Objetivos

O objetivo geral deste Trabalho de Conclusão de Curso é avaliar comparativamente o desempenho de três modelos de Inteligência Artificial — Gemini, DeepSeek e ChatGPT — na identificação e caracterização de e-mails de *phishing*, considerando métricas como precisão, *recall*, acurácia e F1-score.

Para alcançar esse objetivo, foram definidos os seguintes objetivos específicos:

- Identificação de um *dataset* com as amostras representativas de e-mails maliciosos (*phishing*) para análise;
- Identificar as características-chave (*red flags*) presentes em e-mails de *phishing*;
- Avaliar e comparar a eficácia dos modelos Gemini, DeepSeek e ChatGPT na identificação, categorização e detecção de e-mails maliciosos, por meio da aplicação do *dataset* selecionado;
- Documentar e divulgar os resultados obtidos, contribuindo para o avanço da pesquisa nas áreas de segurança da informação e Inteligência Artificial.

1.2 Justificativa

A detecção de e-mails de *phishing* representa um desafio significativo na segurança cibernética, uma vez que os ataques desse tipo evoluem constantemente, tornando-se cada vez mais sofisticados e difíceis de identificar. Desde a década de 1990, criminosos cibernéticos têm adaptado suas estratégias, criando e-mails que vão desde erros gramaticais óbvios até mensagens meticulosamente elaboradas, que simulam comunicações legítimas, enganando até usuários mais experientes. Esses ataques, frequentemente direcionados a indivíduos e organizações, têm causado prejuízos financeiros significativos e danos à reputação de empresas, tornando a busca por métodos eficazes de detecção e prevenção de *phishing* de extrema importância ([Trend Micro, 2024](#)).

Simultaneamente, o uso crescente de Modelos de Linguagem Grande (LLMs) tem mostrado seu potencial para resolver uma série de problemas complexos, incluindo a análise de textos e o reconhecimento de padrões. De acordo com dados de tráfego da web das ferramentas globais de IA, divulgados pela [AITools \(2024\)](#), mais de 10.500 ferramentas de inteligência artificial receberam, juntas, mais de 101 bilhões de visitas ao longo de 2024. Nesse contexto, o Brasil ocupa a terceira posição no ranking global de acessos a essas ferramentas.

O ChatGPT o primeiro lugar, registrou 5,2 bilhões de visitas em fevereiro de 2025, enquanto o Gemini teve 139,4 milhões. O DeepSeek teve seu destaque no início desse ano e teve uma ascensão meteórica em popularidade. Atualmente, é o segundo *chatbot* mais popular, com mais de 792,6 milhões de visitas totais em fevereiro de 2025 ([Android Authority, 2025](#)).

Neste contexto, este estudo justifica-se pelo objetivo de investigar a eficácia dos LLMs na identificação de e-mails de *phishing*, focando na identificação das características-chave, ou *red flags*, que são tipicamente associadas a essas mensagens maliciosas. A definição clara dessas características permitirá não apenas uma análise mais estruturada e objetiva da capacidade de detecção dos modelos, mas também oferecerá parâmetros importantes para a avaliação de ferramentas de segurança por profissionais e usuários finais.

A pesquisa também propõe a testar amostras selecionadas com os modelos Gemini, DeepSeek e ChatGPT, avaliando a eficácia desses modelos na detecção e categorização de *phishing*. A análise comparativa, utilizando métricas como precisão, *recall*, acurácia e F1-score, permitirá identificar qual desses modelos apresenta a maior eficácia na identificação das *red flags* presentes em e-mails maliciosos. A inclusão do DeepSeek, como um modelo mais recente, agrega valor ao estudo, permitindo avaliar tecnologias emergentes frente às ameaças cibernéticas em constante evolução. Destaca-se que, até o momento da realização deste estudo que foi conduzido entre dezembro de 2024 e abril de 2025, não

foram encontrados trabalhos relacionados que utilizassem o DeepSeek especificamente na detecção de e-mails de *phishing*, o que reforça a originalidade e relevância da proposta.

Os resultados obtidos podem fornecer fundamentos essenciais para profissionais de segurança da informação na escolha das ferramentas mais eficazes para a proteção contra ataques de *phishing*, contribuindo para o fortalecimento das estratégias de cibersegurança e ampliando o conhecimento sobre a aplicação de LLMs nesse campo.

2 Fundamentação Teórica

Este capítulo tem como objetivo apresentar os conceitos fundamentais que são essenciais para a compreensão deste trabalho, incluindo literaturas relevantes e contextos específicos que auxiliarão no entendimento do estudo.

2.1 Conceitos Básicos

Esta seção apresenta os principais conceitos que fundamentam este estudo, com foco em ataques de *phishing*, técnicas de inteligência artificial e os modelos de linguagem utilizados. A compreensão desses elementos é essencial para a contextualização da proposta.

2.1.1 Phishing e Engenharia Social

O *phishing* é um tipo de ataque cibernético baseado em engenharia social, que visa enganar usuários para que forneçam informações sensíveis ou realizem ações que comprometam sua segurança digital. Esses ataques podem ocorrer por meio de e-mails, mensagens ou sites fraudulentos, que simulam comunicações legítimas (IBM, 2024; FORTINET, 2024).

A engenharia social, por sua vez, explora a confiança humana para manipular comportamentos. Ataques desse tipo não dependem de falhas técnicas, mas da indução psicológica das vítimas, podendo ocorrer por qualquer canal de comunicação (GOV, 2021).

2.1.2 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial (IA) é uma área da ciência da computação dedicada ao desenvolvimento de sistemas capazes de simular comportamentos inteligentes, como tomada de decisão, raciocínio e aprendizado (TOTVS, 2025). Um dos principais subconjuntos da IA é o *Machine Learning* (ML), que permite a aprendizagem automática por meio de dados históricos.

O ML possibilita que algoritmos ajustem seu comportamento com base em padrões encontrados nos dados, dispensando a programação explícita para cada tarefa. Técnicas como redes neurais e aprendizado profundo têm ampliado significativamente as aplicações do ML em segurança cibernética, particularmente em tarefas de classificação, como a detecção de e-mails maliciosos (Google Cloud, 2024).

2.1.3 Modelos de IA Generativa e LLMs

A IA generativa representa uma subárea recente da IA focada na geração autônoma de conteúdo, como textos, imagens, códigos e vídeos, a partir de entradas fornecidas pelos usuários, os chamados *prompts* (CLOUDFLARE, 2025a). Essa tecnologia tem sido amplamente adotada em soluções de atendimento, geração de conteúdo e, mais recentemente, em segurança da informação.

Dentro desse escopo, os LLMs se destacam por serem treinados com enormes volumes de dados e por empregarem redes neurais profundas. Isso permite que esses modelos compreendam, interpretem e gerem linguagem natural com alto nível de sofisticação (IBM, 2023; CLOUDFLARE, 2025b).

2.1.4 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um ramo da IA responsável por permitir que máquinas compreendam e manipulem linguagem humana em forma escrita ou falada (ORACLE, 2021). Combinado ao aprendizado de máquina, o PLN viabiliza sistemas capazes de extrair significado semântico de textos, sendo uma das tecnologias fundamentais por trás dos LLMs.

2.1.5 Ferramentas baseadas em LLMs

Diversos LLMs têm sido disponibilizados comercialmente por grandes empresas de tecnologia, permitindo interações via *chatbots*. Entre os modelos principais, estão:

- **ChatGPT:** Desenvolvido pela OpenAI, é um dos modelos mais populares, baseado nas arquiteturas GPT-3.5 e GPT-4, com capacidades robustas de geração textual e suporte a instruções complexas (ChatGPT FAQ, 2025);
- **Gemini:** Criado pelo Google DeepMind (anteriormente chamado Bard), é um modelo multimodal que integra texto, imagem, áudio e vídeo em uma única plataforma, com aplicações diversas que vão além do *chat* (Blog do Google Brasil, 2023);
- **DeepSeek:** Desenvolvida pela DeepSeek Company, especializada em modelos de linguagem LLM para processamento e geração de texto, análise de documentos e auxílio em tarefas técnicas, com suporte a arquivos e contexto extenso (DeepSeek, 2025).

Essas ferramentas representam diferentes abordagens no desenvolvimento de LLMs e oferecem oportunidades distintas para aplicações em cibersegurança.

2.2 Trabalhos Relacionados

Atualmente, a literatura apresenta diversas abordagens para a detecção de e-mails de *phishing*, cada uma com metodologias, técnicas e resultados distintos. O estudo de [Rosa, Gringoli e Bellicini \(2024\)](#) propõe um sistema que combina análise estática (examinando cabeçalhos, corpo do texto, links e anexos) e análise dinâmica (utilizando o modelo para interpretação e classificação). Testado com 2.000 e-mails, o sistema alcançou 75,75% de precisão geral, com impressionante sensibilidade de 98,4% na detecção de *phishing*, embora com especificidade moderada de 53,1% para e-mails legítimos. Ao ser comparado com a ferramenta FortiSandbox em 1.000 amostras, o modelo ChatGPT (gpt-3.5-turbo-16k) apresentou correspondência em 49% dos casos. Destacou-se na análise de anexos (72,92%) em comparação à de URLs (46,13%), demonstrando potencial como ferramenta complementar. No entanto, o estudo relatou limitações, como restrições das APIs utilizadas, incapacidade de analisar 57,1% dos componentes dos e-mails, elevado tempo de processamento e dependência de um conjunto de dados restrito, o que compromete sua aplicabilidade em ambientes reais.

No artigo de [Zhang, Jain e Singh \(2024\)](#), o estudo examina o papel duplo do ChatGPT na cibersegurança: como possível vetor de ataque e como mecanismo de defesa. Uma ferramenta automatizada baseada em Python, integrada à API do ChatGPT, foi desenvolvida para aprimorar a detecção de *phishing*. Utilizando 11.430 URLs e 87 características, o modelo GPT-3.5-turbo-0125 demonstrou alta precisão na análise textual. Contudo, dificuldades surgiram na detecção de e-mails em formato de imagem, devido a falhas de reconhecimento visual e problemas de rede. Além disso, observou-se instabilidade após múltiplas iterações (a partir do 25º input), bem como limitações na distinção de mensagens ambíguas e na extração de conteúdo de URLs. O estudo também alerta para os desafios decorrentes da evolução constante dos modelos da OpenAI, dificultando sua adoção estável em contextos de segurança.

Com uma abordagem mais flexível, [Beydemir et al. \(2024\)](#) propõe um sistema baseado em modelos da família GPT (3.5, 4 e uma versão *fine-tuned*), com análise contextual e categorização de ameaças em seis dimensões, como urgência e manipulação emocional. O sistema, testado com os conjuntos Enron e dados sintéticos, alcançou acurácia de 95%, precisão de 94% e *recall* de 96% com o modelo ajustado. Um módulo Python foi utilizado para ofuscar dados sensíveis, garantindo maior privacidade. Embora promissor, o sistema enfrenta limitações: dependência de dados específicos de treinamento, risco de falsos positivos em e-mails com jargões técnicos, necessidade de APIs externas (impactando desempenho e privacidade), alto custo computacional e menor eficácia frente a ataques sofisticados e recentes.

Explorando tecnologias emergentes, o trabalho de [Daoud et al. \(2024\)](#), avalia modelos como GPT-4o, Gemini 1.5 e Claude 3 na detecção de *phishing* em e-mails e golpes

em redes sociais. Quatro abordagens são comparadas: *zero-shot learning* com LLMs, extração de características com *Random Forest*, *fine-tuning* de SLMs e métodos de *ensemble*. Os LLMs obtiveram alta precisão (até 99,21%), seguidos por modelos ajustados e *ensembles* (98,50% de acurácia). Os dados analisados incluíram e-mails em diversos idiomas e postagens em árabe. Apesar dos resultados expressivos, destacam-se limitações como viés linguístico, custos operacionais elevados, falta de explicabilidade nas decisões dos modelos, e vulnerabilidade a novas táticas de ataque – desafios críticos para aplicações práticas.

No estudo de Sayyafzadeh et al. (2024) os autores combinam modelos de linguagem (LLMs) com análise de sentimentos utilizando o ChatGPT-4 e a ferramenta VADER (*Valence Aware Dictionary and Sentiment Reasoner*). O sistema atingiu 92% de acurácia, superando métodos tradicionais ao considerar nuances emocionais como urgência e medo. A pontuação afetiva dos textos variava de -4 a +4, permitindo uma análise mais refinada do conteúdo malicioso. Contudo, a proposta enfrenta desafios como o alto custo dos LLMs, necessidade de adaptação para línguas menos comuns e implicações éticas relacionadas à privacidade. Apesar disso, apresenta grande potencial para aplicação em filtros corporativos e mecanismos de proteção avançados.

A pesquisa de Chataut, Gyawali e Usman (2024) compara o desempenho de modelos como GPT-3.5, GPT-4 e CyberGPT em um conjunto de 828 e-mails (504 legítimos e 324 *phishing*). Os modelos GPT-4 e CyberGPT obtiveram acurácia superior a 97%, superando o GPT-3.5 (80,68%). O CyberGPT demonstrou maior confiabilidade (94,31%) na detecção de indicadores como remetentes suspeitos e linguagem manipulativa. No entanto, limitações foram observadas: dependência de dados de treinamento, alto custo de modelos *premium*, vulnerabilidade a falsos negativos, dificuldades em contextos ambíguos, necessidade de constante atualização e possíveis latências em tempo real.

Em (HEIDING et al., 2024) os autores avaliaram os modelos GPT-4, Claude-1, Bard (PaLM 2) e LLaMA-2. Claude-1 destacou-se com 100% de acerto em e-mails sofisticados e 75% em genéricos. O GPT-4 teve desempenho intermediário, enquanto o Bard apresentou variabilidade entre 0% e 60%, e o LLaMA-2 detectou apenas 25% a 50% dos casos. O estudo também abordou o uso de LLMs na criação de *phishing*, constatando que métodos manuais ainda são mais eficazes (com taxa de cliques de 69–79%, contra 30–44% do GPT-4). Além disso, os modelos demonstraram alta dependência de *prompts*, variabilidade nos resultados, falta de explicabilidade e suscetibilidade a falsos positivos e negativos, indicando que ainda não substituem especialistas humanos nem garantem proteção completa.

O artigo de Jiang (2024) investigam o uso de GPT-3.5 e GPT-4 na detecção de golpes cibernéticos via análise de padrões linguísticos e contexto textual. O processo envolveu coleta, pré-processamento e rotulagem manual de dados, seguido de *fine-tuning*. Embora os resultados iniciais indiquem eficácia na identificação de indícios como erros gramaticais

e links suspeitos, o estudo limita-se a um único tipo de ataque, sem análises quantitativas robustas ou cenários operacionais. A ausência de testes práticos e considerações sobre escalabilidade restringe a aplicabilidade imediata dos achados.

Em constância com as abordagens anteriores, o trabalho de [Xu, Wang e Jajodia \(2014\)](#) apresenta uma extensão de navegador “anti-*phishing*” que bloqueia sites fraudulentos quando os usuários inserem seus nomes de usuário em domínios não associados às suas contas legítimas. Diferentemente dos métodos tradicionais, o sistema Gemini utiliza o par nome de usuário, domínio para detectar *phishing* com alta precisão, atingindo 0% de falsos negativos e menos de 1% de falsos positivos em testes com 50 sites legítimos e 50 fraudulentos. Implementado como extensão para Chrome, Firefox e Internet Explorer, o Gemini apresenta sobrecarga mínima (<10% no tempo de carregamento) e compatibilidade com outras ferramentas de proteção. Entre suas limitações, destacam-se a incapacidade de proteger nomes de usuário (atuando apenas sobre senhas) e a dificuldade em lidar com páginas de login não padronizadas (como páginas em Flash). Apesar disso, os resultados indicam sua eficácia como uma solução complementar, leve e transparente para a prevenção de vazamento de credenciais em tempo real, com potencial de integração em navegadores e iniciativas como a APWG.

A pesquisa de [Nagarajan e Kamalbabu \(2024\)](#) investiga os riscos associados ao uso de modelos de IA como o ChatGPT e o Google Bard (Gemini) na cibersegurança, com ênfase em sua exploração em ataques de *phishing*, espionagem e disseminação de informações falsas. O estudo recomenda regulamentações rigorosas para coibir usos maliciosos, treinamento dos modelos para detectar fraudes, ações educativas voltadas aos usuários e colaboração internacional. No entanto, apresenta fragilidades, como a ausência de dados empíricos que comprovem o uso prático desses modelos em ataques reais e a carência de evidências quantitativas sobre o aumento efetivo nos riscos de *phishing*. Também se observam limitações nas próprias capacidades dos modelos em prevenir ataques e uma análise pouco aprofundada quanto à implementação das medidas propostas em escala global.

Já em ([BEGOU et al., 2023](#)), explora-se como ferramentas de IA podem ser utilizadas para criar páginas de *phishing* de maneira rápida e acessível, mesmo por usuários sem conhecimento técnico. A IA demonstrou capacidade de gerar páginas visualmente semelhantes às originais, aumentando o potencial de sucesso dos ataques. O estudo também destaca a colaboração entre humanos e máquinas, ressaltando que humanos ainda conseguem identificar nuances que os modelos não captam. No entanto, limitações importantes foram identificadas, como o pequeno tamanho da amostra, que restringe a generalização dos resultados, ausência de testes em cenários reais, dependência de serviços externos e dificuldades técnicas como o limite de *tokens* dos modelos e a baixa eficácia na ofuscação de dados. Além disso, o modelo de aprendizado de máquina utilizado mostrou-se básico, incapaz de capturar as complexidades das estratégias de *phishing* modernas, reforçando

a necessidade de avanços em compreensão semântica no PLN.

O trabalho de [Ramprasath et al. \(2023\)](#) aborda a detecção de ameaças por meio do uso de Redes Neurais Recorrentes (RNNs) com células LSTM, com o objetivo de identificar e-mails de spam e *phishing*. Essa abordagem permite capturar padrões temporais e contextuais em mensagens com comprimentos variados, destacando-se pela robustez a ruídos, adaptabilidade a novos dados e desempenho superior em relação a métodos como SVM e k-NN. Apesar do potencial, a técnica apresenta limitações: alto custo computacional, sensibilidade a variações linguísticas, perda de desempenho em mensagens mal formatadas, com erros ortográficos ou uso de técnicas de ofuscação, além da necessidade de ajuste fino de hiperparâmetros e da baixa interpretabilidade dos modelos.

A análise de [Zhou et al. \(2025\)](#) investiga os riscos ocultos nos grandes modelos de raciocínio, evidenciando sua vulnerabilidade a ataques como *jailbreak* e injeção de *prompt*, além da facilidade de uso para fins maliciosos. Embora tais modelos apresentem maior capacidade de raciocínio e respostas mais detalhadas, esse aprimoramento pode aumentar o risco de geração de respostas prejudiciais. Como solução, os autores propõem técnicas avançadas de treinamento e reforço do raciocínio explícito voltado à segurança. Contudo, o estudo carece de comparações abrangentes com modelos como DeepSeek-R1 e o o3-mini, e não aprofunda as implicações éticas e sociais, tampouco propõe medidas concretas para mitigar os riscos não visíveis no processo de inferência.

O estudo de [Wang, Tang e He \(2025\)](#) explora a capacidade de LLMs de simular comportamentos humanos e dinâmicas sociais complexas, com aplicações em áreas como mercados financeiros, interações culturais e psicologia. Os modelos demonstraram eficiência e escalabilidade, mas enfrentam desafios como a falta de personalização nos dados de treinamento e a presença de distorções que comprometem a fidelidade das simulações. Para mitigar essas limitações, os autores sugerem o enriquecimento dos dados com variáveis psicológicas e experiências pessoais, o aprimoramento dos agentes simulados, a inserção de valores e conhecimento social, e o desenvolvimento de métricas de avaliação mais robustas que considerem aspectos qualitativos e quantitativos.

Por fim, [Rahman et al. \(2025\)](#) realiza uma análise comparativa entre os modelos DeepSeek, ChatGPT e Google Gemini, abordando suas respectivas forças, limitações e direções futuras. O DeepSeek se destaca pela eficiência e especialização em domínios como medicina e direito, embora careça de suporte multimodal e criatividade. O ChatGPT apresenta versatilidade e desempenho sólido em tarefas como redação, pesquisa e atendimento, mas sofre com possíveis imprecisões e desatualização de conhecimento. O Gemini, por sua vez, sobressai na capacidade multimodal e integração ao ecossistema Google, sendo capaz de processar texto, código, imagens, áudio e vídeo. No entanto, enfrenta limitações quanto ao alto consumo de recursos, o que afeta sua acessibilidade, e desafios éticos relacionados à geração de *deepfakes* e informações falsas.

A Tabela 1 demonstra a comparação dos trabalhos relacionados, inclusive considerando o presente trabalho. A seguir, são fornecidas as explicações sobre o significado de cada coluna da tabela.

- **Referência:** Apresenta os trabalhos relacionados que foram comparados com o presente estudo.
- **Escopo:** Descreve o contexto ou tipo de dados analisados nos estudos, como *Phishing (e-mail)* ou *Phishing (rede social)*, indicando onde as técnicas ou modelos foram aplicados.
- **Modelos Avaliados:** Indica os modelos utilizados em cada trabalho, listando as versões ou variantes dos LLMs avaliados, como GPT (por exemplo, GPT-3.5, GPT-4), Gemini, Claude, DeepSeek, LLaMA. A tabela mostra quais modelos foram utilizados para a tarefa de detecção de *phishing*.
- **Caracterização de *Phishing*:** Esta coluna indica se o estudo inclui a caracterização ou extração das características-chave (*red flags*) dos e-mails de *phishing*. Um (✓) significa que a caracterização foi feita. No caso deste estudo, a caracterização de *phishing* é uma das contribuições metodológicas principais, sendo que os modelos analisados no presente trabalho realizam essa tarefa.
- **Dataset de Características de *Phishing*:** Indica se o estudo gerou um novo conjunto de dados com base nas análises realizadas. Um (✓) representa a criação de um *dataset* expandido com as respostas dos modelos. Já um (–) indica que o trabalho utilizou apenas *datasets* existentes, sem modificações ou complementações. Este estudo se destaca por construir um *dataset* derivado, enriquecido com características-chave (*red flags*) extraídas automaticamente de e-mails de *phishing*, baseado nas análises dos LLMs realizadas.

O artigo de [Heiding et al. \(2024\)](#) não especifica a versão exata do modelo Gemini (Bard) utilizada. O símbolo (✓) foi empregado para indicar que o modelo foi de fato utilizado na pesquisa. O símbolo (–) indica que o modelo ou característica não foi abordada.

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

Referência	Escopo	Modelos Avaliados					Caracterização de Phishing	Dataset de Características de Phishing
		GPT	Gemini	Claude	DeepSeek	LLaMA		
Rosa, Gringoli e Bellicini (2024)	Phishing (e-mail)	3.5	–	–	–	–	–	–
Zhang, Jain e Singh (2024)	Phishing (e-mail)	3.5 / 4	–	–	–	–	–	–
Beydemir et al. (2024)	Phishing (e-mail)	4 + FT	–	–	–	–	–	–
Daoud et al. (2024)	Phishing (e-mail, rede social)	4	1.5	3	–	–	–	–
Sayyafzadeh et al. (2024)	Phishing (e-mail)	4	–	–	–	–	–	–
Chataut, Gyawali e Usman (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	–	✓	✓

3 Desenvolvimento

Este estudo foi conduzido em etapas distintas, abrangendo a seleção do *dataset*, a padronização e organização dos dados, a caracterização dos e-mails, sua classificação e posterior análise. A seguir, são descritas em detalhe as principais fases da metodologia adotada.

O processo metodológico adotado para a obtenção dos resultados da análise proposta está ilustrado na Figura 1 e será detalhado nas seções a seguir.

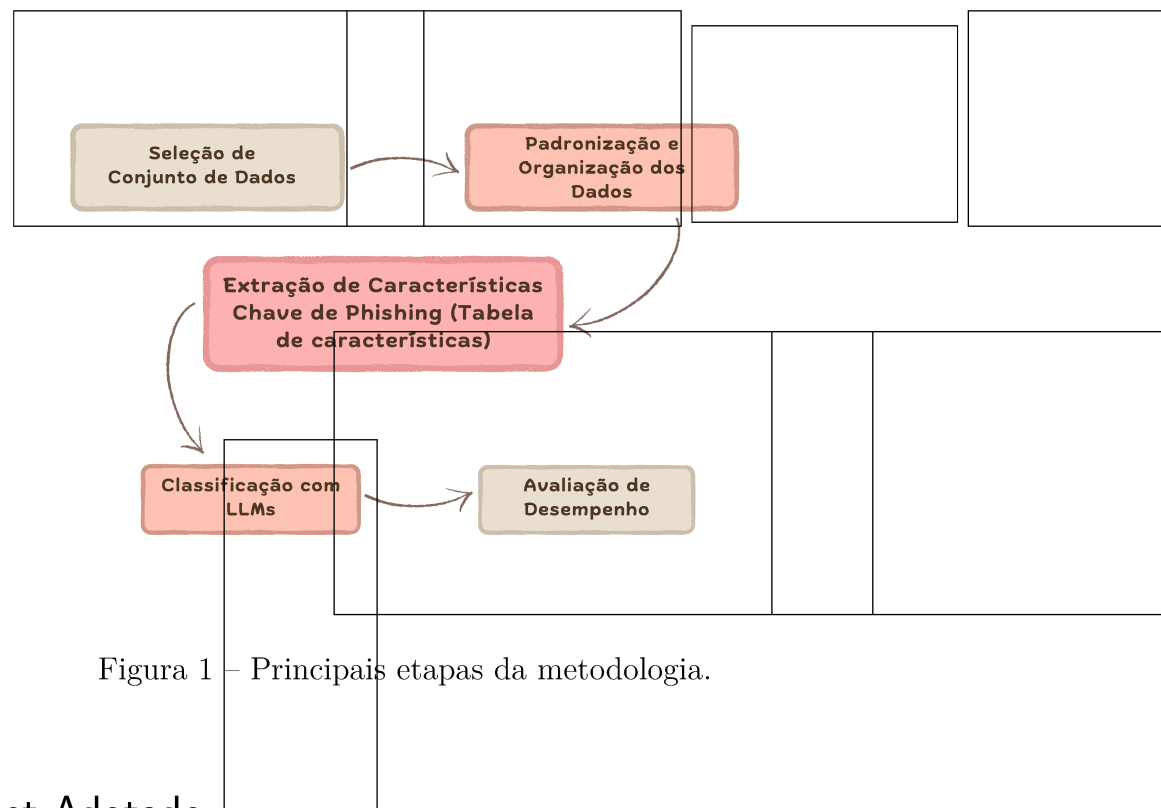


Figura 1 – Principais etapas da metodologia.

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 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.

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

3.2 Organização dos dados

Foram utilizadas 1.009 amostras, número definido com base na média observada em trabalhos relacionados, o que permite uma comparação adequada dos resultados. Desse total, 602 correspondem a e-mails legítimos (*Safe E-mail*) e 407 a e-mails maliciosos (*Phishing E-mail*).

3.3 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 seguintes comandos:

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, solicitando aos modelos que identificassem características ainda não catalogadas. No entanto, observamos significativa redundância nos novos atributos identificados, o que nos levou a remoção das características repetidas. Este processo resultou em conjuntos otimizados e diferenciados de características para cada LLM, os quais serão detalhadamente apresentados e discutidos nas subseções seguintes, destacando as particularidades da abordagem analítica de cada modelo.

3.3.1 Ajustes Finais

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*

- **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 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*.

3.3.2 Gemini

Após o refinamento realizado com o Gemini, as características identificadas e suas respectivas descrições estão apresentadas nas Tabelas 2 e 3.

#	Característica	Descrição
1	Remetente suspeito	Verifique se o endereço de e-mail do remetente é legítimo. Muitas vezes, os golpistas usam endereços semelhantes aos de empresas conhecidas, mas com pequenas variações.
2	Saudação genérica	E-mails legítimos geralmente usam seu nome. Se a saudação for “Prezado cliente” ou algo semelhante, desconfie.
3	E-mail não solicitado	Se você não se inscreveu para receber o e-mail, é um sinal de alerta.
4	Senso de urgência/medo	Golpistas frequentemente criam um senso de urgência para que você aja rapidamente, sem pensar.
5	Oferta boa demais	Se a oferta parecer irreal, provavelmente é.
6	Conteúdo impróprio	E-mails com conteúdo sexualmente explícito ou ofensivo são geralmente <i>phishing</i> .
7	Solicitação de informações confidenciais	Empresas legítimas nunca solicitam senhas ou informações bancárias por e-mail.
8	<i>Links</i> suspeitos	Passe o mouse sobre os <i>links</i> para verificar o endereço real. Se o endereço for diferente do esperado, não clique.
9	Anexos suspeitos	Não abra anexos de remetentes desconhecidos.
10	Remetente falsificado	É possível falsificar o remetente do e-mail, então mesmo que o e-mail pareça ser de um local confiável, ainda sim deve-se ter cautela.

Tabela 2 – Características comuns de e-mails de *phishing* detectados pelo Gemini (Parte 1 de 2).

#	Característica	Descrição
11	Domínio suspeito	Verifique se o domínio do e-mail confere com o domínio da empresa que supostamente enviou o e-mail.
12	Histórias elaboradas	Muitas vezes e-mails de <i>phishing</i> contam histórias complexas para tentar te enganar.
13	Personalização excessiva	Às vezes o e-mail pode conter muitas informações pessoais, o que pode passar uma sensação de segurança, mas é justamente aí que mora o perigo.
14	Falta de informação de contato	E-mails de empresas confiáveis sempre possuem informações de contato.
15	Linguagem inadequada	Erros gramaticais e ortográficos são comuns em e-mails de <i>phishing</i> .
16	Conteúdo emocionalmente apelativo	Golpistas podem usar suas emoções para manipular você.
17	Informações de contato inconsistentes	Verifique se as informações de contato no e-mail correspondem às informações no site da empresa.
18	Formatação estranha	E-mails de <i>phishing</i> podem ter formatação inconsistente ou imagens de baixa qualidade.
19	Anexos com nomes estranhos	Anexos com nomes genéricos ou estranhos podem conter <i>malware</i> .
20	Probabilidade de <i>Phishing</i>	Quanto mais sinais de alerta um e-mail tiver, maior a probabilidade de ser <i>phishing</i> .

Tabela 3 – Características comuns de e-mails de *phishing* detectados pelo Gemini (Parte 2 de 2).

3.3.3 DeepSeek

A Tabela 4 mostra as características identificadas pelo modelo DeepSeek.

#	Característica	Descrição
1	Assunto alarmante ou urgente	Assuntos que criam uma sensação de urgência.
2	Remetente desconhecido ou suspeito	Remetentes que você não reconhece ou que parecem falsos.
3	<i>Links</i> suspeitos	<i>Links</i> que não correspondem ao site oficial da empresa ou que contêm erros ortográficos.
4	Solicitação de informações pessoais	Pedidos de informações confidenciais, como senhas ou números de cartão de crédito.
5	Erros gramaticais e ortográficos	Textos com muitos erros de gramática e ortografia.
6	Promoções ou ofertas irreais	Ofertas que parecem boas demais para ser verdade.
7	Falta de personalização	Mensagens genéricas que não usam seu nome ou informações específicas.
8	Uso de domínios falsos ou similares	Domínios que imitam o nome de empresas conhecidas, mas com pequenas alterações.
9	Botões ou <i>links</i> enganosos	Botões que parecem levar a um site legítimo, mas redirecionam para um site falso.
10	Uso excessivo de urgência	Mensagens que pressionam o destinatário a agir imediatamente.
11	Probabilidade de <i>Phishing</i>	Classificação da probabilidade de <i>phishing</i> .

Tabela 4 – Características comuns de e-mails de *phishing* detectado o DeepSeek.

3.3.4 ChatGPT

Concluindo, a Tabela 5 mostra as características identificadas pelo modelo ChatGPT.

#	Característica	Descrição
1	Remetente Suspeito	Indica se o remetente é desconhecido, tem um nome genérico ou parece forjado (ex: e-mails vindos de domínios gratuitos ou com nomes estranhos).
2	Assunto Alarmante	Verifica se o assunto do e-mail tenta gerar urgência, medo ou curiosidade exagerada (ex: “Sua conta será encerrada!” ou “Ganhe dinheiro agora!”).
3	Erros Gramaticais	Detecta se o texto contém muitos erros de ortografia ou gramática — algo comum em e-mails de <i>phishing</i> mal feitos.
4	<i>Links</i> Suspeitos	Indica se há <i>links</i> encurtados, URLs estranhas ou diferentes do site legítimo (ex: domínios parecidos com os verdadeiros, como “g00gle.com”).
5	Anexos Inesperados	Verifica se há anexos que não foram solicitados e que podem conter malware (ex: arquivos .exe, .zip, .scr, etc.).
6	Pedido de Informações Sensíveis	Checa se o e-mail solicita dados como senhas, número de cartão de crédito ou CPF.
7	Saudação Genérica	Indica se o e-mail usa uma saudação genérica como “Prezado cliente” em vez de nomear o destinatário — típico de e-mails automáticos de <i>phishing</i> .
8	Formato Incomum	Observa se o e-mail tem estrutura visual estranha, má formatação, excesso de espaços, fontes irregulares ou estilo inconsistente.
9	E-mail Não Solicitado	Indica se o e-mail foi recebido sem o usuário ter se inscrito ou solicitado, geralmente vindo de campanhas em massa.
10	Endereço de Resposta Diferente	Verifica se o endereço de resposta (“Reply-To”) é diferente do endereço de envio, o que pode indicar tentativa de redirecionar a comunicação para um fraudador.
11	Probabilidade de <i>Phishing</i>	Classificação baseada na soma dos fatores acima.

Tabela 5 – Características comuns de e-mails de *phishing* detectado o ChatGPT.

3.4 Avaliação e Comparação de Resultados

A fim de avaliar e comparar o desempenho dos modelos, foram utilizadas as métricas de Precisão, *Recall*, Acurácia e F1-Score, amplamente adotadas na literatura por avaliarem diferentes aspectos do desempenho de modelos de classificação. Essas métricas permitem uma análise mais robusta, evitando vieses e capturando múltiplas dimensões do desempenho dos modelos.

Precisão: Proporção de previsões positivas que estavam corretas.

Fórmula:

$$\text{Precisão} = \frac{TP}{TP + FP}$$

***Recall*:** Proporção de casos positivos reais que foram corretamente identificados.

Fórmula:

$$\text{recall} = \frac{TP}{TP + FN}$$

Acurácia (*Accuracy*): Proporção de previsões corretas (tanto positivas quanto negativas) em relação ao total.

Fórmula:

$$\text{Acurácia} = \frac{TP + TN}{TP + FP + FN + TN}$$

F1-Score: Média entre precisão e *recall*.

Fórmula:

$$\text{F1 Score} = 2 \times \frac{\text{Precisão} \times \text{recall}}{\text{Precisão} + \text{recall}}$$

As siglas utilizadas nas equações representam os seguintes significados:

TP: Positivos corretamente identificados.

TN: Negativos corretamente identificados.

FP: Positivos classificados de forma errada (falsos alarmes).

FN: Negativos classificados de forma errada (falsas omissões).

Essas métricas são fundamentais para a avaliação de modelos de classificação, especialmente no contexto da detecção de e-mails de *phishing*, em que tanto falsos positivos quanto falsos negativos podem gerar impactos significativos.

3.5 Dataset Gerado

Como contribuição metodológica adicional deste estudo, desenvolvemos um *dataset* derivado que amplia substancialmente o potencial analítico da investigação. Este conjunto de dados vai além dos e-mails originais testados, incluindo também as características de *phishing* que cada modelo conseguiu identificar. Para cada e-mail analisado, registramos tanto os resultados da classificação quanto as métricas de desempenho, o que nos permite entender melhor o comportamento dos modelos.

Com este recurso, conseguimos comparar os diferentes métodos de detecção de maneira clara, identificando onde cada abordagem se destaca e onde apresenta dificuldades. Disponibilizamos o *dataset* completo no Apêndice A, tornando o trabalho transparente e permitindo que outros pesquisadores possam reproduzir e expandir os experimentos com facilidade.

4 Resultados

Neste capítulo, são apresentados os resultados obtidos a partir da aplicação da metodologia descrita no capítulo anterior. Vale ressaltar que, nos gráficos apresentados, o termo *performance* refere-se ao desempenho dos modelos de linguagem nas métricas de avaliação utilizadas: precisão, *recall*, acurácia e F1-Score.

4.1 Gemini

Durante as análises dos resultados, observou-se que, na maioria dos casos, o Gemini classificava corretamente a coluna “Probabilidade de *Phishing*” conforme os critérios estabelecidos (Baixa, Média ou Alta). No entanto, foi identificado que, especialmente em e-mails mais curtos, o modelo retornava o valor “Nenhuma” como resposta, divergindo do padrão esperado.

Essa divergência impactou os resultados da análise. Para garantir consistência, os e-mails classificados como “nenhuma” foram submetidos a novos testes, utilizando os mesmos dados. Em alguns casos, o modelo continuou retornando “nenhuma”. Diante disso, adotou-se o critério de substituir esse valor por “Baixa”. Essa adaptação resultou em uma melhoria perceptível nos indicadores de desempenho. A Figura 2 apresenta os resultados das métricas precisão, *recall*, acurácia e F1-Score, considerando as amostras classificadas como “nenhuma”.

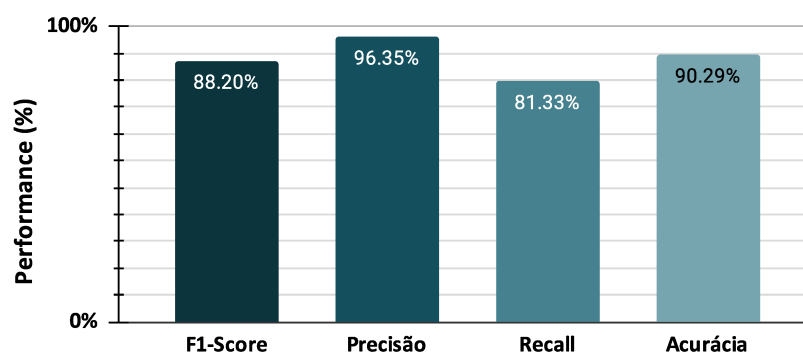


Figura 2 – Resultados com o Gemini considerando o valor “nenhuma”, na coluna “Probabilidade de Phishing”

A atribuição do valor “nenhuma” pode indicar uma limitação do modelo em interpretar mensagens muito curtas, o que pode comprometer a confiabilidade da classificação. Isso levanta uma possível falha de análise por parte do Gemini, que deve ser considerada nas interpretações dos resultados.

Na Figura 2, observam-se os seguintes valores: F1-Score de 88,20%, precisão de 96,35%, *recall* de 81,33% e acurácia de 90,29%. A Tabela 6 apresenta a matriz de confusão referente a essa primeira análise.

Valores Reais	Valores Previstos	
	Positivo	Negativo
Positivo	331	76
Negativo	22	580

Tabela 6 – Matriz de Confusão do Gemini considerando o valor “nenhuma”, na coluna “Probabilidade de Phishing”.

Após a substituição do valor “nenhuma” por “Baixa”, os dados foram reprocessados. A Figura 3 apresenta os resultados atualizados das métricas após essa alteração, evidenciando melhorias no desempenho geral do modelo.

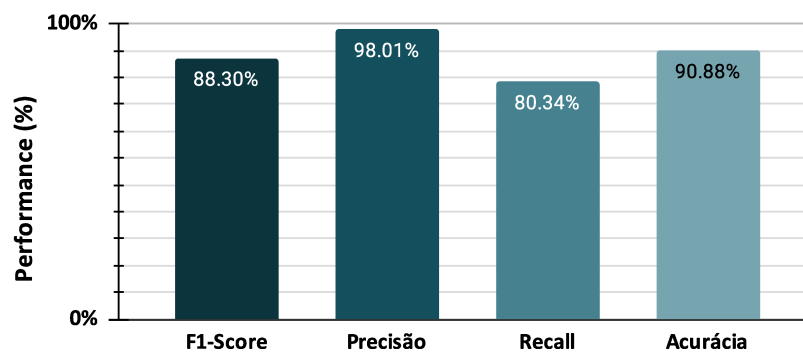


Figura 3 – Resultados com o Gemini após a alteração das amostras classificadas como “nenhuma” para “baixa”, na coluna “Probabilidade de Phishing”

Os resultados apresentados na Figura 3 demonstram um desempenho superior ao da Figura 2 em todas as métricas, com exceção do *recall*. Foram obtidos F1-Score de 88,30%, precisão de 98,01%, *recall* de 80,34% e acurácia de 90,88%. Esses valores indicam uma melhoria na capacidade de classificação após a substituição do valor “nenhuma” por “baixa” na coluna “Probabilidade de Phishing”.

A Tabela 7 apresenta a matriz de confusão do Gemini após essa alteração.

Valores Reais	Valores Previstos	
	Positivo	Negativo
Positivo	327	80
Negativo	12	590

Tabela 7 – Matriz de Confusão do Gemini após a alteração das amostras classificadas como “nenhuma” para “baixa”, na coluna “Probabilidade de Phishing”.

4.2 DeepSeek

A Figura 4 apresenta os resultados do modelo DeepSeek, com F1-Score de 92,38%, precisão de 96,68%, *recall* de 88,45% e acurácia de 93,36%. Esses valores indicam um desempenho superior ao do modelo Gemini (Figura 3) em todas as métricas, com exceção da precisão. A Tabela 8 exibe a matriz de confusão correspondente aos resultados do DeepSeek.

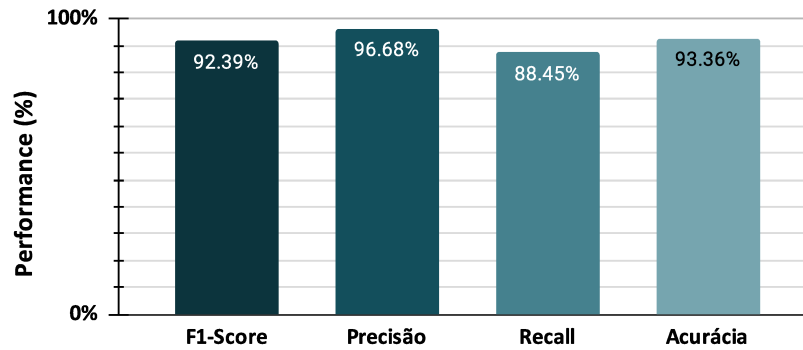


Figura 4 – Resultados do DeepSeek

Valores Reais	Valores Previstos	
	Positivo	Negativo
Positivo	360	47
Negativo	20	582

Tabela 8 – Matriz de Confusão do DeepSeek

4.3 ChatGPT

A Figura 5 apresenta os resultados das métricas obtidos pelo ChatGPT, que alcançou um F1-Score de 89,23%, precisão de 90,05%, *recall* de 88,42% e acurácia de 89,40%. Em comparação com os resultados do DeepSeek, o ChatGPT apresentou valores inferiores em todas as métricas. Já em relação ao Gemini, o ChatGPT obteve desempenho superior nas métricas de F1-Score e *recall*. A Tabela 9 mostra a matriz de confusão dos resultados do DeepSeek.

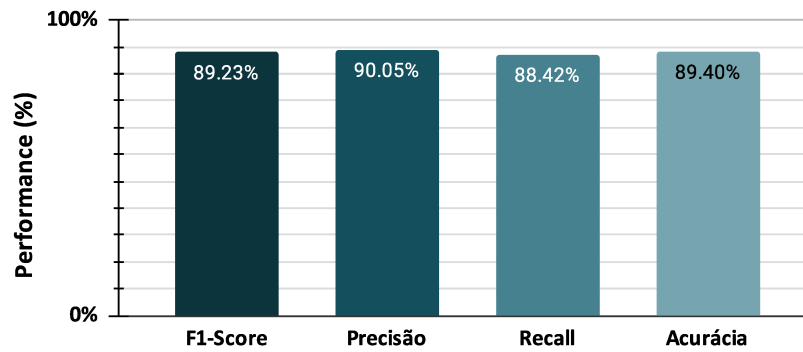


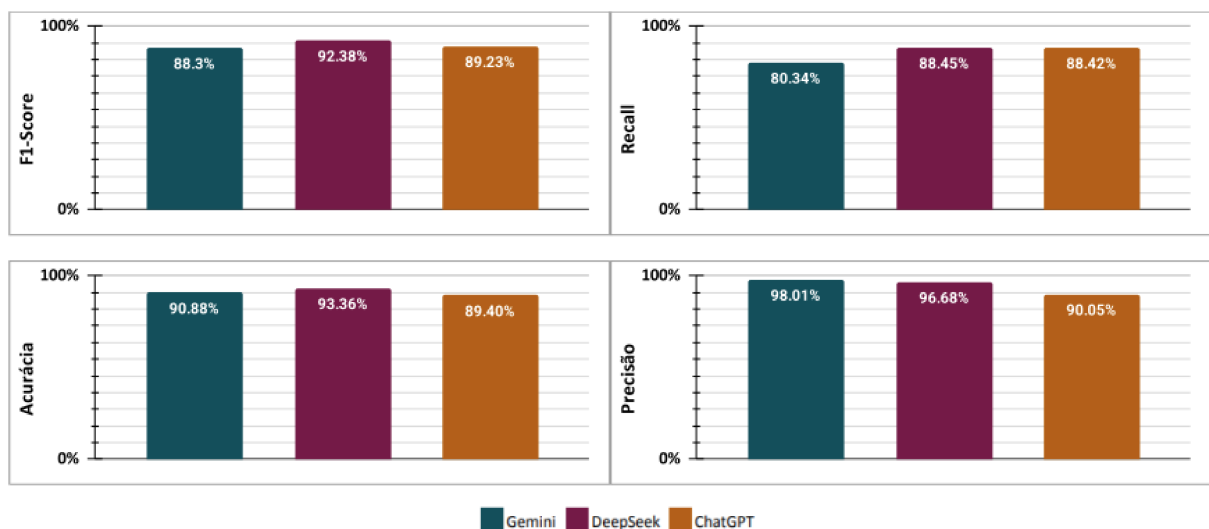
Figura 5 – Resultados do ChatGPT

Valores Reais	Valores Previstos	
	Positivo	Negativo
Positivo	359	47
Negativo	60	543

Tabela 9 – Matriz de Confusão do ChatGPT

4.4 Comparações

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 6. 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.

Figura 6 – Resultados consolidados das métricas de desempenho dos modelos Gemini, DeepSeek e ChatGPT: Precisão, *recall*, Acurácia 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 fundamental considerar quais são os requisitos prioritários da organização ou do usuário final do sistema de e-mails.

5 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*.

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.

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.

Referências

- AITOOLS. **Global AI Adoption Statistics Trends**. 2024. Acessado em 10 maio de 2025. Disponível em: <<https://aitools.xyz/statistics/2024/>>. Citado na página 13.
- Android Authority. **Estatísticas mostram que o uso do ChatGPT está deixando Gemini para trás, e nem chega perto**. 2025. Acessado em 10 maio de 2025. Disponível em: <<https://www.androidauthority.com/gemini-trailing-competitors-3540687/>>. Citado na página 13.
- Associação APWG. **Relatórios de tendências de atividades de phishing**. 2025. Acessado em 09 de abril de 2025. Disponível em: <<https://apwg.org/trendsreports/>>. Citado na página 11.
- BEGOU, N.; VINOY, J.; DUDA, A.; KORCZYŃSKI, M. Exploring the dark side of ai: Advanced phishing attack design and deployment using chatgpt. In: **2023 IEEE Conference on Communications and Network Security (CNS)**. [S.l.: s.n.], 2023. p. 1–6. Citado na página 19.
- BEYDEMIR, A. B.; SEZGIN, U.; DOĞAN, U.; AŞIKLAR, B. E.; YERLIKAYA, F. A.; BAHTIYAR, A. A dynamically selected gpt model for phishing detection. In: **2024 14th International Conference on Advanced Computer Information Technologies (ACIT)**. [S.l.: s.n.], 2024. p. 481–484. Citado 3 vezes nas páginas 11, 17 e 22.
- Blog do Google Brasil. **Apresentando o Gemini: nosso maior e mais hábil modelo de IA**. 2023. Acessado em 03 de abril de 2025. Disponível em: <<https://blog.google/intl/pt-br/novidades/tecnologia/apresentando-o-gemini-nosso-maior-e-mais-habil-modelo-de-ia/#mensagem-sundar>>. Citado na página 16.
- CHAKRABORTY, S. **Phishing Email Detection**. Kaggle, 2023. Disponível em: <<https://www.kaggle.com/dsv/6090437>>. Citado na página 23.
- CHATAUT, R.; GYAWALI, P. K.; USMAN, Y. 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)**. [S.l.: s.n.], 2024. p. 0548–0554. Citado 3 vezes nas páginas 11, 18 e 22.
- ChatGPT FAQ. **O que é o ChatGPT?** 2025. Acessado em 03 de abril de 2025. Disponível em: <<https://chatgpt.com.br/faq/#o-que-e-chatgpt>>. Citado na página 16.
- CISCO. **O que é phishing?** 2024. Acessado em 01 maio de 2025. Disponível em: <[https://www.cisoadvisor.com.br/](https://www-cisco-com.translate.goog/site/us/en/learn/topics/security/what-is-phishing.html?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=sge#:~:text=Phishing%C3%A9%20pr%C3%A1tica%20de,malware%20no%20dispositivo%20da%20v%C3%ADtima.> https://www-cisco-com.translate.goog/site/us/en/learn/topics/security/what-is-phishing.html?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=sge#:~:text=Phishing%C3%A9%20pr%C3%A1tica%20de,malware%20no%20dispositivo%20da%20v%C3%ADtima.>>. Citado na página 11.</p>
<p>Ciso Advisor. E-mail de phishing gerado por IA é quase impossível de detectar. 2023. Acessado em 01 maio de 2025. Disponível em: <

[e-mail-de-phishing-gerado-por-ia-e-quase-impossivel-de-detectar/>](#). Citado na página 11.

CLOUDFLARE. **O que é IA generativa?** 2025. Acessado em 02 de abril de 2025. Disponível em: [<https://www.cloudflare.com/pt-br/learning/ai/what-is-large-language-model/>](https://www.cloudflare.com/pt-br/learning/ai/what-is-large-language-model/>). Citado na página 16.

_____. **O que é um modelo de linguagem ampla (LLM)?** 2025. Acessado em 03 de abril de 2025. Disponível em: [<https://www.cloudflare.com/pt-br/learning/ai/what-is-large-language-model/>](https://www.cloudflare.com/pt-br/learning/ai/what-is-large-language-model/>). Citado na página 16.

CNN Brasil. **Brasil é vice-campeão em ataques cibernéticos, com 1.379 golpes por minuto, aponta estudo.** 2024. [<https://www.cnnbrasil.com.br/economia/negocios/brasil-e-vice-campeao-em-ataques-ciberneticos-com-1-379-golpes-por-minuto-aponta-estudo/>](https://www.cnnbrasil.com.br/economia/negocios/brasil-e-vice-campeao-em-ataques-ciberneticos-com-1-379-golpes-por-minuto-aponta-estudo/>). Acesso em: 28 março. 2025. Citado na página 11.

DAOUD, E. A.; DAOUD, L. A.; ASASSFEH, M.; AL-SHAIKH, A.; AL-SHERIDEH, A. S.; AFANEH, S. Enhancing cybersecurity with transformers: Preventing phishing emails and social media scams. In: **2024 IEEE Conference on Dependable and Secure Computing (DSC)**. [S.l.: s.n.], 2024. p. 31–36. Citado 3 vezes nas páginas 11, 17 e 22.

DeepSeek. **Atualização do modelo DeepSeek-V3, avanço abrangente de vários recursos.** 2025. Acessado em 01 de maio de 2025. Disponível em: [<https://api-docs.deepseek.com/zh-cn/news/news250325>](https://api-docs.deepseek.com/zh-cn/news/news250325>). Citado na página 16.

FORTINET. **19 tipos de ataques de phishing.** 2024. Acessado em 01 de abril de 2025. Disponível em: [<https://www.fortinet.com/br/resources/cyberglossary/types-of-phishing-attacks>](https://www.fortinet.com/br/resources/cyberglossary/types-of-phishing-attacks>). Citado na página 15.

Google Cloud. **O que é machine learning (ML)?** 2024. Acessado em 03 de abril de 2025. Disponível em: [<https://cloud.google.com/learn/what-is-machine-learning?hl=pt-BR>](https://cloud.google.com/learn/what-is-machine-learning?hl=pt-BR>). Citado na página 15.

GOV, P. **Engenharia social. Guia para Proteção de Conhecimentos Sensíveis.** 2021. Acessado em 01 de abril de 2025. Disponível em: [<https://www.gov.br/abin/pt-br/institucional/acoes-e-programas/PNPC/boaspraticas/cartilha-engenharia-social-guia-para-protecao-de-conhecimentos-sensiveis>](https://www.gov.br/abin/pt-br/institucional/acoes-e-programas/PNPC/boaspraticas/cartilha-engenharia-social-guia-para-protecao-de-conhecimentos-sensiveis>). Citado na página 15.

HEIDING, F.; SCHNEIER, B.; VISHWANATH, A.; BERNSTEIN, J.; PARK, P. S. Devising and detecting phishing emails using large language models. **IEEE Access**, v. 12, p. 42131–42146, 2024. Citado 4 vezes nas páginas 11, 18, 21 e 22.

IBM, I. B. M. C. **O que é LLM (grandes modelos de linguagem)?** 2023. Acessado em 01 de abril de 2025. Disponível em: [<https://www.ibm.com/br-pt/think/topics/large-language-models>](https://www.ibm.com/br-pt/think/topics/large-language-models>). Citado na página 16.

_____. **O que é phishing?** 2024. Acessado em 01 de abril de 2025. Disponível em: [<https://www.ibm.com/br-pt/topics/phishing>](https://www.ibm.com/br-pt/topics/phishing>). Citado na página 15.

JIANG, L. **Detecting Scams Using Large Language Models**. 2024. Disponível em: <<https://arxiv.org/abs/2402.03147>>. Citado 3 vezes nas páginas 11, 18 e 22.

MEDIUM. **Modelos de Grandes Linguagens (LLMs) em Segurança Cibernética: Uma Mudança de Paradigma na Inteligência de Ameaças**.

2024. Acessado em 01 maio de 2025. Disponível em: <<https://mawgoud.medium.com/large-language-models-llms-in-cybersecurity-a-paradigm-shift-in-threat-intelligence-5e0f7653dc11>>. Citado na página 11.

NAGARAJAN, S.; KAMALBABU, K. Analysis of how chatgpt and gemini help in the generation of cyber attacks. In: **2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)**. [S.l.: s.n.], 2024. p. 663–667. Citado na página 19.

ORACLE. **O que é Processamento de Linguagem Natural (PLN)?**

2021. Acessado em 03 de abril de 2025. Disponível em: <[https://www.oracle.com/br/artificial-intelligence/what-is-natural-language-processing/#:~:text=O%20processamento%20de%20linguagem%20natural%20\(NLP\)%20%C3%A9%20um%20ramo%20da,ou%20voz%20de%20linguagem%20natural.](https://www.oracle.com/br/artificial-intelligence/what-is-natural-language-processing/#:~:text=O%20processamento%20de%20linguagem%20natural%20(NLP)%20%C3%A9%20um%20ramo%20da,ou%20voz%20de%20linguagem%20natural.)> Citado na página 16.

RAHMAN, A.; MAHIR, S. H.; TASHRIF, M. T. A.; AISHI, A. A.; KARIM, M. A.; KUNDU, D.; DEBNATH, T.; MOUDUDI, M. A. A.; EIDMUM, M. Z. A. **Comparative Analysis Based on DeepSeek, ChatGPT, and Google Gemini: Features, Techniques, Performance, Future Prospects**. 2025. Disponível em: <<https://arxiv.org/abs/2503.04783>>. Citado na página 20.

RAMPRASATH, J.; PRIYANKA, S.; MANUDEV, R.; GOKUL, M. Identification and mitigation of phishing email attacks using deep learning. In: **2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)**. [S.l.: s.n.], 2023. p. 466–470. Citado na página 20.

ROSA, S. de; GRINGOLI, F.; BELLICINI, G. Hey chatgpt, is this message phishing? In: **2024 22nd Mediterranean Communication and Computer Networking Conference (MedComNet)**. [S.l.: s.n.], 2024. p. 1–10. Citado 3 vezes nas páginas 11, 17 e 22.

SAYYAFZADEH, S.; WEATHERSPOON, M.; YAN, J.; CHI, H. 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)**. [S.l.: s.n.], 2024. p. 159–165. Citado 3 vezes nas páginas 11, 18 e 22.

TOTVS. **O guia completo e atualizado sobre Inteligência Artificial**. 2025.

Acessado em 01 de abril de 2025. Disponível em: <<https://www.totvs.com/blog/inovacoes/o-que-e-inteligencia-artificial/>>. Citado na página 15.

Trend Micro. **Quais são os diferentes tipos de ataques de phishing?** 2024.

Acessado em 01 maio de 2025. Disponível em: <https://www.trendmicro.com/pt_br/what-is/phishing/types-of-phishing.html#:~:text=O%20phishing%20de%20e%2Dmail,e%2Dmail%20que%20possam%20obter.> Citado na página 13.

WANG, Q.; TANG, Z.; HE, B. **From ChatGPT to DeepSeek: Can LLMs Simulate Humanity?** 2025. Disponível em: <<https://arxiv.org/abs/2502.18210>>. Citado na página 20.

XU, Z.; WANG, H.; JAJODIA, S. Gemini: An emergency line of defense against phishing attacks. In: **2014 IEEE 33rd International Symposium on Reliable Distributed Systems**. [S.l.: s.n.], 2014. p. 11–20. Citado na página 19.

ZHANG, D.; JAIN, K.; SINGH, P. Guarding against chatgpt threats: Identifying and addressing vulnerabilities. In: **2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)**. [S.l.: s.n.], 2024. p. 612–615. Citado 3 vezes nas páginas 11, 17 e 22.

ZHOU, K.; LIU, C.; ZHAO, X.; JANGAM, S.; SRINIVASA, J.; LIU, G.; SONG, D.; WANG, X. E. **The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1**. 2025. Disponível em: <<https://arxiv.org/abs/2502.12659>>. Citado na página 20.

Apêndices

APÊNDICE A – Acesso ao Dataset

O dataset completo utilizado nesta pesquisa está disponível no seguinte endereço:

<https://github.com/evelyn-bustamante/Phishing_Analise_Resultados>