Detecção de Mensagens de Phishing Geradas por Modelos Generativos de Linguagem em Português por Meio de Atributos Estilométricos Resilientes a Vieses de Geração Sintética

Jeanderson Medeiros da Silva[1], **Jaqueline Damacena Duarte**[1], **Fabio Lucio Lopes de Mendonca**[1], **Laerte Peotta de Melo**[1]

[1] Departamento de Engenharia Elétrica
Universidade de Brasília (UnB) – Brasília, DF – Brasil

{jeanderson.silva, jaqueline.duarte}@aluno.unb.br
fabio.mendonca@redes.unb.br, laerte.melo@unb.br

Abstract. *This article investigates the potential of a predictive model based on behavioral stylometry to identify AI-generated phishing in Portuguese. The methodology focuses on extracting morphological attributes, such as imperative commands, to mitigate biases in synthetic models. Given the scenario of dataset shift between architectures (Gemini and GPT-4o), an XGBoost classifier was evaluated, achieving an F1-Score of 0.91 in a controlled environment. The results indicate that stylometric analysis offers relevant clues to complement traditional defenses, although generalization in real-world scenarios still requires validation with organic data.*

Resumo. *Este artigo investiga o potencial de um modelo preditivo baseado em estilometria comportamental para identificar phishing gerado por IA em Português. A metodologia foca na extração de atributos morfológicos, como comandos imperativos, para mitigar vieses de modelos sintéticos. Diante do cenário de dataset shift entre arquiteturas (Gemini e GPT-4o), avaliou-se um classificador XGBoost que alcançou F1-Score de 0,91 em ambiente controlado. Os resultados indicam que a análise estilométrica oferece indícios relevantes para complementar defesas tradicionais, embora a generalização em cenários reais ainda demande validação com dados orgânicos.*

1. Introdução

Ataques de engenharia social, particularmente o *phishing*, continuam a representar uma das ameaças mais críticas à segurança da informação global. Com a popularização de Modelos de Linguagem de Grande Escala (LLMs) como ChatGPT e Gemini, atacantes ganharam a capacidade de gerar mensagens em massa com fluência gramatical, eliminando erros ortográficos clássicos que tradicionalmente serviam como gatilhos para filtros de *spam* [Heiding et al. 2024]. Essa evolução tecnológica resultou em ameaças que emulam o comportamento humano de forma consideravelmente persuasiva. A natureza dinâmica e polimórfica das táticas contemporâneas de *phishing* torna os paradigmas tradicionais de defesa, como heurísticas baseadas em assinaturas e *blacklists* estáticas, obsoletos e ineficazes contra ataques de dia zero (*zero-day*) [Saeed 2025]. Para suplantar

essa vulnerabilidade, a literatura recente aponta que a solução reside no aprendizado de máquina, especificamente no uso de modelos de conjunto (*ensembles*), que combinam múltiplas inteligências preditivas para criar barreiras adaptáveis contra ameaças emergentes [Saeed 2025].

Para mitigar esse cenário, a literatura recente tem explorado a detecção estilométrica, que analisa as características intrínsecas e linguísticas do texto, como estrutura sintática e classes gramaticais [Opara et al. 2025]. No entanto, a pesquisa nesta área esbarra em um obstáculo significativo: a escassez de *datasets* públicos de e-mails de *phishing* reais gerados por IA. Consequentemente, os modelos de detecção dependem de *datasets* sintéticos, o que frequentemente induz algoritmos de aprendizado de máquina ao *overfitting*. Nesses casos, os modelos memorizam atalhos preditivos da IA, como a extensão padrão do texto gerado ou jargões promocionais, em vez de aprenderem a intenção maliciosa subjacente [Eze and Shamir 2024].

Este artigo busca explorar essas limitações ao investigar a resiliência da detecção estilométrica de *phishing* em língua portuguesa. O estudo apresenta uma abordagem metodológica que visa expandir a análise para além da detecção em ambientes estritamente controlados, analisando falhas de generalização em cenários adversariais e testando alternativas que auxiliem na mitigação de vieses sintéticos de geração.

Diante desse cenário, e considerando a escassez de *datasets* públicos representativos de ameaças reais geradas por IA, este estudo estrutura-se como uma prova de conceito conduzida em ambiente controlado. Ainda assim, busca-se assegurar rigor analítico por meio da simulação de condições adversariais relevantes. Nesse contexto, a principal contribuição do trabalho reside na adaptação e validação de atributos estilométricos para o idioma Português, com foco específico em:

- A extração customizada do atributo F41 (comandos persuasivos), que expande a captura de verbos imperativos para construções do modo subjuntivo com intenção de ação;
- A criação de um corpus adversarial em Português (Modo Red Team) projetado para isolar o viés dimensional (extensão do texto);
- Uma análise experimental do *dataset shift* entre diferentes LLMs, demonstrando como a diversidade de arquiteturas geradoras afeta a resiliência de classificadores clássicos.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 detalha a metodologia, incluindo a geração dos *datasets* adversariais e a extração dos atributos estilométricos adaptados para o português. A Seção 4 discute os resultados experimentais, o estudo de ablação e a validação do modelo híbrido. Por fim, a Seção 5 apresenta a conclusão e os trabalhos futuros.

2. Trabalhos Relacionados

A literatura recente sobre segurança da informação tem se ramificado para acompanhar a evolução das ameaças digitais. Para contextualizar a contribuição deste estudo, os trabalhos correlatos foram divididos em três eixos principais: abordagens clássicas baseadas em conteúdo textual, o impacto da inteligência artificial na geração de ameaças e o uso da estilometria como mecanismo de defesa.

2.1. Detecção de Phishing Baseada em Conteúdo Textual

Tradicionalmente, a identificação de e-mails maliciosos e sites de *phishing* dependia de listas de bloqueio (*blacklists*) e da extração de características de URLs, como o comprimento do domínio e a presença de caracteres especiais [Tang and Mahmoud 2021, Patil et al. 2022]. Contudo, à medida que os atacantes começaram a utilizar infraestruturas legítimas para hospedar links maliciosos, a comunidade científica passou a focar no processamento do corpo do texto [Bauskar et al. 2024].

Pesquisas contemporâneas empregam Processamento de Linguagem Natural (PLN) para extrair vetores baseados na frequência de termos e jargões suspeitos [Bauskar et al. 2024]. Esses vetores alimentam algoritmos clássicos de aprendizado de máquina, como *Random Forest* e *Support Vector Machines* (SVM), que demonstram alta eficácia na separação de mensagens legítimas e maliciosas baseando-se no vocabulário e estilo adotado pelo atacante [Alhaji et al. 2025, Opara et al. 2025]. No entanto, esses métodos heurísticos assumem que o atacante cometerá erros ortográficos clássicos ou usará um formato rígido e previsível, premissa que deixou de ser uma constante com a evolução das ferramentas generativas [Heiding et al. 2024].

2.2. Detecção de Textos Produzidos por Inteligência Artificial

A popularização de Modelos de Linguagem de Grande Escala (LLMs) introduziu um novo paradigma na engenharia social [Heiding et al. 2024]. Atacantes agora utilizam ferramentas generativas para produzir e-mails de *phishing* em massa, eliminando os erros gramaticais que historicamente serviam como os principais gatilhos de detecção para os filtros de *spam* [Eze and Shamir 2024].

Estudos recentes investigam como diferenciar o texto humano do texto sintético. O estudo de [Eze and Shamir 2024], por exemplo, contrastou mensagens de *phishing* geradas por IA com o *corpus* humano da Enron, demonstrando que a IA produz padrões textuais únicos e uma sintaxe muitas vezes mais polida que a humana. [Opara et al. 2025] expandiram esse conceito provando que os LLMs podem ser instruídos a modular o nível de sofisticação do ataque, utilizando gatilhos psicológicos como urgência ou oportunidades financeiras de forma altamente persuasiva e adaptável.

2.3. Estilometria Aplicada a Tarefas de Segurança

Para contornar a capacidade da IA de emular o texto humano, a estilometria, a análise quantitativa do estilo de escrita, tem se destacado. Em vez de buscar palavras-chave maliciosas, essa técnica extrai dezenas de marcadores estruturais, como densidade de pronomes, frequências de pontuação, uso de conjunções e complexidade de leitura [Opara et al. 2025]. A premissa é que, independentemente do tema, a intenção persuasiva de comando deixará ”impressões digitais”na estrutura sintática do texto.

A aplicação da estilometria, contudo, é consideravelmente sensível ao idioma. As fórmulas tradicionais de legibilidade, como o *Flesch Reading Ease*, foram construídas para a morfologia da língua inglesa, exigindo adaptações matemáticas rigorosas para refletir adequadamente as características estruturais e o esforço cognitivo no português brasileiro [Martins et al. 1996].

Embora a literatura supracitada valide a eficácia da estilometria [Opara et al. 2025] e documente o alto nível de ameaça do *phishing* sintético

[Eze and Shamir 2024], observa-se uma lacuna: a dependência de *datasets* padronizados que induzem os modelos a memorizarem a extensão do texto em vez de sua intenção, além da escassez de adaptações metodológicas para a língua portuguesa.

Ademais, embora o estado da arte contemporâneo incorpore *baselines* modernos baseados em *sentence embeddings* e arquiteturas profundas de *Transformers* (e.g., BERTimbau, mBERT e XLM-R) para a classificação de textos, a elevada acurácia desses modelos frequentemente está associada a menor transparência decisória, sendo comumente caracterizados como "caixas-pretas" (*black-boxes*) [Opara et al. 2025, Owa and Adewole 2025]. Em cenários envolvendo ameaças sintéticas, compreender *como* a manipulação sintática pode contribuir para evasão de detecção mostra-se relevante, além do próprio bloqueio da mensagem. Nesse contexto, a literatura tem destacado a importância de modelos com maior interpretabilidade (*white-box*) [Opara et al. 2025, Patil et al. 2025]. Este trabalho busca preencher essa lacuna ao investigar o uso de métodos de ablação para mitigar o *overfitting* dimensional e avalia a resiliência do modelo frente ao *dataset shift*, por meio de validação cruzada em um *dataset* híbrido.

3. Metodologia

3.1. Arquitetura Geral do Framework

A arquitetura do *framework* proposto foi estruturada de forma modular, com o intuito de facilitar a reprodutibilidade e a auditoria de cada etapa do processamento de dados. Inspirada em metodologias de detecção baseadas em aprendizado de máquina [Opara et al. 2025], a abordagem integra a geração primária de dados adversariais a uma etapa de extração de atributos estilométricos, com adaptações voltadas à morfologia da língua portuguesa.

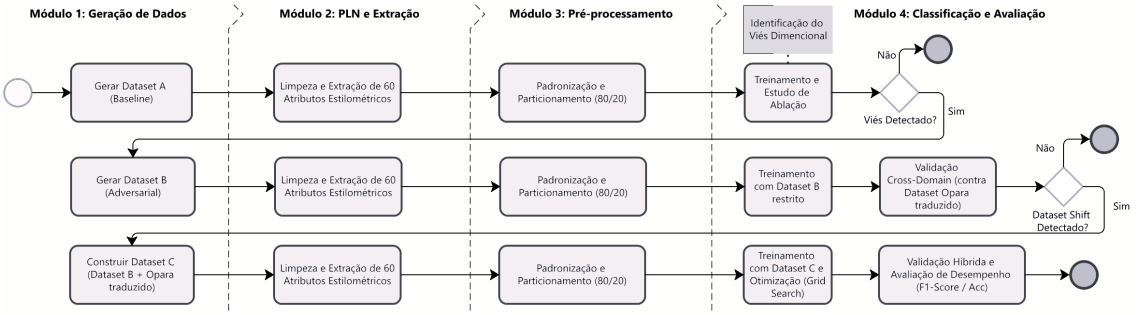

Figura 1. Fluxograma da arquitetura metodológica, abrangendo desde a geração dos dados adversariais até a otimização híbrida do classificador.

Conforme ilustrado na Figura 1, o *pipeline* operacional do *framework* estrutura-se em quatro módulos sequenciais: (1) Geração de Dados, que interage com o LLM *Gemini* sob restrições de *prompt* (modo *Red Team*) para sintetizar um *corpus* de e-mails de *phishing* e comunicações legítimas; (2) Processamento de Linguagem Natural (PLN), que aplica rotinas de tokenização e *POS Tagging* via biblioteca *spaCy* (modelo `pt_core_news_lg`) para extrair e quantificar 60 atributos estilométricos; (3) Pré-processamento e Transformação, que emprega a técnica *StandardScaler* para normalizar os vetores, impedindo que atributos com amplitudes desproporcionais dominem a função

de perda durante o treinamento; e (4) Classificação e Avaliação Híbrida, no qual o algoritmo *XGBoost* é otimizado via *Grid Search* e validação cruzada para avaliar sua resiliência contra o *dataset shift* [Opara et al. 2025]. Essa arquitetura modular viabiliza o controle e isolamento de variáveis (e.g., estudos de ablação), forçando o classificador a priorizar as assinaturas linguísticas de persuasão e atenuando a memorização de ruídos estruturais específicos das IAs.

3.2. Engenharia de *Prompt*, Higienização e Construção dos *Datasets*

A geração do *corpus* sintético utilizou a API do modelo *Gemini-2.5-flash*, com os parâmetros de moderação (*safety settings*) desativados (`BLOCK_NONE`) para viabilizar a simulação de ataques de engenharia social. Visando refletir ameaças reais e persuasivas, a geração adotou uma taxonomia de sete temas de alta prevalência em relatórios globais de inteligência cibernética (APWG, Cofense e Verizon) [Opara et al. 2025]: (1) ofertas promocionais; (2) oportunidades financeiras; (3) emprego; (4) atualizações de TI; (5) saúde; (6) notificações financeiras; e (7) relacionamentos.

A estruturação dos dados foi executada em três fases independentes, visando isolar variáveis e garantir a resiliência do classificador. A Tabela 1 sumariza a composição, a fonte e o objetivo metodológico de cada conjunto gerado.

Tabela 1. Resumo da composição e objetivo dos *datasets* estruturados para o experimento.

Dataset	Distribuição	Fonte / Gerador	Objetivo Metodológico
A (Baseline)	150 instâncias (75 *phishing*, 75 legítimos)	*Gemini* (Padrão)	Avaliar o comportamento orgânico da IA e identificar atalhos preditivos.
B (Evasivo)	500 instâncias (250 *phishing*, 250 legítimos)	*Gemini* (Modo *Red Team*)	Mitigar vieses dimensionais por meio da restrição de comprimento e vocabulário.
C (Híbrido)	626 instâncias (*Dataset* B + 126 amostras de [Opara et al. 2025]	*Gemini* + *GPT-4o*)	Atestar a capacidade de generalização do modelo contra o *dataset shift*.

A seguir, detalham-se as restrições e as diretrizes de *prompt* aplicadas na construção de cada conjunto de dados, cujas transcrições literais, referentes à geração dos *datasets* A e B, encontram-se no Apêndice B.

- ***Dataset* A (Baseline):** Os *prompts* instruíram o modelo a modular a sofisticação do ataque (básico, intermediário e avançado), explorando gatilhos psicológicos comuns e variações dinâmicas de tema e ação.
- ***Dataset* B (Red Team / Evasivo):** Para mitigar os vieses dimensionais identificados na fase exploratória, aplicaram-se regras restritivas estritas. O modelo foi forçado a gerar textos contendo entre 80 e 100 palavras, sendo explicitamente proibido o uso de jargões de urgência ou termos promocionais, emulando comunicações corporativas furtivas. Para fomentar a reprodutibilidade da pesquisa, o *Dataset* B encontra-se publicamente acessível[1].

[1] `https://github.com/jeanderson-mdrs/AI-Generated-Phish-Detector.git`

- ***Dataset* C (Híbrido):** Constitui a mescla do Dataset B com 126 amostras externas originárias do estudo de [Opara et al. 2025](63 *phishing* e 63 legítimas), cujo *dataset* também é público e acessível no Kaggle. As instâncias foram traduzidas para o português via *DeepL*. A literatura atesta a superioridade empírica desta arquitetura neural em fluência e adequação semântica [Silva and Pires 2024], mitigando a introdução de ruídos sintáticos artificiais que poderiam corromper a extração estilométrica.

Para assegurar a integridade da extração de características textuais, os dados brutos foram submetidos a uma rotina de higienização, etapa indispensável no pré-processamento textual [Tang and Mahmoud 2021]. Artefatos sistemáticos de metadados (e.g., "Assunto:" e "Corpo:") foram removidos, fundamentando-se no princípio da Frequência Documental Inversa (IDF), que postula que termos onipresentes não agregam poder discriminativo ao modelo [Barik et al. 2025, Tang and Mahmoud 2021]. Para mitigar o risco de *data leakage*, também foram expurgados *placeholders* artificiais inseridos pelas IAs para adjetivar URLs (e.g., *link* "malicioso", "fictício", "falso" ou "disfarçado") e formatações em *markdown* (como asteriscos duplos), cuja aglutinação corrompia a tokenização morfológica. Instâncias vazias ou com falhas de API foram excluídas. Por fim, para evitar a criação de diferenças artificiais entre as classes, a geração do *Dataset* B seguiu restrições simétricas: mensagens maliciosas e legítimas foram forçadas à mesma dimensão (80 a 100 palavras), isentas de *feedbacks* da IA e orientadas a adotar um tom corporativo formal. Esse procedimento contribui para que o classificador aprenda a intenção semântica, e não vieses sintéticos de formatação.

3.3. Estrutura Computacional e Extração de Atributos Estilométricos

A etapa de engenharia de atributos foi implementada por meio de um *script* desenvolvido na linguagem Python, com suporte das bibliotecas de Processamento de Linguagem Natural *spaCy* (utilizando o modelo de precisão `pt_core_news_lg`) e *Textstat*. Para fins de auditoria e reprodutibilidade, o *Dataset* B encontra-se publicamente acessível no repositório supracitado. O processamento de cada instância do *dataset* consistiu na segmentação do texto em *tokens*, análise de dependência sintática e anotação morfológica (*POS Tagging*), resultando na extração de um vetor de 60 atributos estilométricos (F1 a F60) por e-mail [Opara et al. 2025].

A adoção dessa quantidade específica de variáveis fundamentou-se em [Opara et al. 2025], escolhido como trabalho-base por estabelecer o estado da arte ao capturar nuances sintáticas intrínsecas a IAs geradoras. Além de compilarem métricas clássicas eficazes, os autores provaram empiricamente que LLMs modulam a sofisticação de ataques usando gatilhos psicológicos adaptáveis, tornando seus preditores ideais para validar a resiliência do modelo frente ao *dataset shift*. Para o presente trabalho, esse vetor de 60 atributos foi transposto e adaptado às regras morfológicas da língua portuguesa, focando na detecção de padrões textuais atípicos e persuasão coercitiva. A Tabela 7, disponível no Apêndice A, detalha essas características, organizadas em oito dimensões analíticas.

Destacam-se duas contribuições metodológicas para a adaptação do processamento à língua portuguesa. Primeiramente, o cálculo de fluência de leitura de *Flesch* (`F31`) sofreu uma adequação matemática. Uma vez que os vocábulos em português pos-

suem, em média, maior extensão silábica que no idioma inglês, aplicou-se a Equação 1 proposta por [Martins et al. 1996]:

$$Flesch_{PT} = 248,835 - (1,015 \times ASL) - (84,6 \times ASW) \quad (1)$$

O ASL representa o tamanho médio da sentença (*Average Sentence Length*) e ASW indica a média de sílabas por palavra (*Average Syllables per Word*). A aplicação isolada da fórmula original norte-americana inflacionaria a complexidade e penalizaria injustamente a legibilidade de comunicações corporativas legítimas em português.

Em segundo lugar, a extração de comandos persuasivos (`F41`), um forte indicador comportamental em e-mails de *phishing*, foi recalibrada sintaticamente. Além de capturar a classe morfológica oficial de imperativo (`Mood=Imp`), o algoritmo foi programado para rastrear verbos do modo subjuntivo (`Mood=Sub`) atrelados a intenções de ação (e.g., ”esperamos que você acesse”, ”para que você verifique”). Esta expansão buscou atenuar a técnica evasiva adotada por IAs geradoras, que tendem a mascarar ameaças sob uma estrutura sintática polida em detrimento do comando imperativo clássico.

3.4. Configuração Experimental e Otimização do Modelo

A modelagem computacional foi conduzida em ambiente Python, valendo-se das bibliotecas *Scikit-Learn* e *XGBoost*. Antes de submeter os dados aos algoritmos de aprendizado, aplicou-se uma etapa de pré-processamento utilizando a técnica de *StandardScaler*. A adoção dessa etapa de normalização é recomendada para favorecer que os modelos sejam treinados com dados de maior qualidade. Essa prática sistemática ajuda a evitar que atributos com escalas de grandeza naturalmente elevadas (e.g., a contagem total de caracteres) exerçam uma influência matemática desproporcional sobre atributos de escala fracionária (e.g., a densidade de pronomes) durante o cálculo vetorial do algoritmo, o que tende a resultar em um desempenho mais estável em aplicações do mundo real [Owa and Adewole 2025].

A fim de garantir a representatividade das classes em todas as fases da pesquisa, a técnica de fracionamento *Hold-out* com amostragem estratificada foi adotada como padrão. Tanto nas avaliações individuais (utilizando os Datasets A e B) quanto na validação do modelo híbrido (Dataset C), os dados foram divididos na proporção de 80% para a base de treinamento e 20% para o conjunto de teste. Para assegurar a reprodutibilidade exata do particionamento dos dados e dos resultados por outros pesquisadores, fixou-se a semente de aleatoriedade (`random_state=42`) em todas as instâncias e algoritmos do experimento. Essa padronização metodológica permitiu a comparação isolada do desempenho em cada nível de restrição adversarial.

Na fase inicial, avaliou-se o desempenho de quatro classificadores base amplamente referenciados na literatura de segurança da informação: Regressão Logística, *Support Vector Machine* (SVM), *Random Forest* e *XGBoost*. Embora o *Random Forest* e o *SVM* tenham apresentado alta precisão preditiva nos *datasets* isolados da prova de conceito, a adoção do *XGBoost* como estimador base para a arquitetura híbrida definitiva não se baseou apenas na performance empírica, mas também nas recomendações consolidadas no estado da arte. [Opara et al. 2025] demonstram que a arquitetura de otimização de erros do *XGBoost* o torna significativamente mais resiliente na modelagem de rela-

cionamentos não lineares intrínsecos à linguagem persuasiva, sendo, portanto, a escolha metodológica ideal para manter a eficácia da detecção em cenários adversariais no mundo real.

Para atenuar os riscos de falha de generalização decorrentes do *dataset shift* (i.e., a diferença estrutural entre e-mails gerados pelos modelos Gemini e GPT), o modelo final foi submetido a um processo sistemático de calibração de hiperparâmetros (*Fine-Tuning*). A otimização foi conduzida por meio do método de *Grid Search* acoplado à técnica de Validação Cruzada Estratificada de 5 partições (*5-Fold Cross-Validation*). O espaço de busca avaliou permutações controladas dos hiperparâmetros: `learning_rate` (0.01, 0.1, 0.2), `max_depth` (4, 6, 8), `n_estimators` (100, 200, 300), `subsample` (0.8, 1.0) e `colsample_bytree` (0.8, 1.0), totalizando 540 ajustes interativos.

Para assegurar a reprodutibilidade integral, os scripts, configurações e as versões das bibliotecas utilizadas (e.g., spaCy e XGBoost) foram disponibilizados publicamente no repositório supracitado.

A configuração com melhor desempenho na validação cruzada convergiu para a arquitetura apresentada na Tabela 2. A adoção de uma profundidade ampliada aliada a um número controlado de estimadores permitiu ao classificador capturar interações estilométricas mais complexas, favorecendo características universais de persuasão (e.g., verbos imperativos) em detrimento de ruídos do *dataset* sintético. Adicionalmente, a amostragem estocástica tanto de instâncias quanto de atributos pulverizou o peso preditivo, atuando como um mecanismo de regularização frente ao *dataset shift* e reduzindo a dependência de uma única arquitetura geradora.

Tabela 2. Configuração otimizada dos hiperparâmetros do modelo *XGBoost*.

Hiperparâmetro	Valor	Justificativa
`n_estimators`	100	Garante convergência estável sem onerar o custo computacional.
`max_depth`	8	Profundidade ampliada para mapear as interações estilométricas complexas no *dataset* híbrido.
`learning_rate`	0,2	Taxa de aprendizado acelerada, equilibrando a profundidade das árvores na correção de erros.
`subsample`	0,8	Subamostragem estocástica (80% das instâncias), forçando o modelo a generalizar.
`colsample_bytree`	0,8	Amostra 80% das características por árvore, pulverizando o peso preditivo e mitigando a dominância de atributos isolados.

3.5. Métricas de Avaliação

O desempenho preditivo do modelo foi avaliado pelas métricas de Acurácia, Precisão, Revocação e *F1-Score*, assumindo o *phishing* sintético como classe positiva [Opara et al. 2025]. Devido à assimetria inerente às ameaças cibernéticas, o *F1-Score* foi adotado como métrica principal de otimização durante o *Grid Search*, visando penalizar vieses de classificação que favoreçam desproporcionalmente uma das classes e assegurar o aprendizado da verdadeira intenção estilométrica da mensagem.

4. Resultados e Discussão

Esta seção detalha os resultados empíricos obtidos durante as fases de treinamento e validação da arquitetura proposta. A análise prossegue desde o diagnóstico de vieses até a validação do modelo em cenários de domínio cruzado.

4.1. Desempenho Inicial e Diagnóstico de Viés Algorítmico

A avaliação preliminar do *framework* consistiu no treinamento e na comparação dos quatro algoritmos de aprendizado de máquina selecionados (Regressão Logística, SVM, *Random Forest* e *XGBoost*), operando sobre o espectro completo de 60 atributos estilométricos obtidos conforme descrito na Seção 3.3, Tabela 7.

Os resultados desta fase inicial indicaram um desempenho preditivo aparentemente ideal. Os modelos não lineares (SVM, *Random Forest* e *XGBoost*) convergiram para métricas perfeitas de Acurácia e *F1-Score* $(1,00)$, enquanto a Regressão Logística alcançou um *F1-Score* de $0,96$.

Embora resultados máximos sejam numericamente atrativos, na literatura de detecção de *phishing* baseado em LLMs, pontuações perfeitas em *datasets* sintéticos frequentemente sinalizam a ocorrência de superajuste (*overfitting*) ou a memorização de artefatos estruturais intrínsecos à IA geradora [Tang and Mahmoud 2021].

Para investigar a resiliência dessa superfície de decisão, procedeu-se à análise da importância dos atributos (*Feature Importance*) do modelo *XGBoost*, ancorada no critério de Ganho de Informação (*Information Gain*). Foi observado um desequilíbrio acentuado na distribuição de importância, indicando uma dependência excessiva com relação a variável F48 (30% do Ganho de Informação).

Essa constatação evidenciou que o algoritmo não estava aprendendo primordialmente a intenção persuasiva e a estilometria sutil da ameaça, mas sim atuando como um simples filtro de palavras-chave, ancorado nos jargões temáticos (e.g., ”oferta”, ”promoção”) contidos nos *prompts* iniciais. Tal diagnóstico confirmou a necessidade de uma depuração dos atributos por meio de um estudo de ablação.

4.2. Estudo de Ablação no Modelo Base (Dataset A)

A ablação é uma técnica analítica que consiste na remoção sistemática de atributos para mensurar o impacto individual de cada variável na superfície de decisão do modelo [Tang and Mahmoud 2021]. O experimento avaliou o modelo *XGBoost* sob quatro restrições progressivas: (A) o modelo completo; (B) a exclusão isolada da variável dominante `F48` (palavras promocionais); (C) a exclusão dos 5 atributos mais informativos; e (D) a exclusão dos 10 atributos principais.

Conforme sumarizado na Tabela 3, a remoção do ”atalho lexical”`F48` causou uma redução imediata na estabilidade preditiva do modelo, com o *F1-Score* decaindo de $1,00$ para $0,92$. Quando o algoritmo foi privado de seus 10 melhores preditores, a métrica estabilizou-se na margem de $0,82$.

A degradação observada evidencia que o classificador dependia substancialmente do vocabulário de *marketing* imposto pelos *prompts* não restritivos do *Dataset A*. Uma análise de importância residual no Cenário D (sem os Top 10) revelou uma reorganização

Tabela 3. Degradação do desempenho preditivo durante o estudo de ablação (Dataset A).

Cenário de Restrição	Atributos Ativos	Acurácia	*F1-Score*
Cenário A: Modelo Completo	60	1,000	1,000
Cenário B: Sem `F48` (Promocional)	59	0,923	0,923
Cenário C: Sem Top 5 Atributos	55	0,807	0,814
Cenário D: Sem Top 10 Atributos	50	0,807	0,827

na árvore de decisão. Privado de seus principais indicadores, o modelo transferiu a prioridade preditiva para atributos estruturais secundários, com destaque para a densidade de pronomes (`F18`) e, significativamente, para a contagem total de caracteres (`F2`), sugerindo uma inclinação ao viés dimensional.

4.3. Análise de Resiliência e Diagnóstico do Viés Dimensional

A ascensão de variáveis de comprimento como mecanismo de predição levantou questionamentos sobre a resiliência do modelo em ambientes não controlados. Para validar essa hipótese, desenvolveu-se uma avaliação adversarial, submetendo o classificador treinado no *Dataset A* a um conjunto de teste isolado composto por 40 e-mails de *phishing* evasivos.

Neste teste, a arquitetura geradora dos dados do *dataset* de teste operou sob instruções em modo *Red Team*, sendo instruída a suprimir intencionalmente qualquer jargão promocional e adotar uma linguagem estritamente corporativa e direta. Ao avaliar essas instâncias evasivas, o modelo inicial manteve uma taxa de detecção nominal de 100% e probabilidade de confiança média de $0,96$.

Contudo, a análise residual sugere que, na ausência de jargões promocionais, o classificador passou a ancorar sua decisão predominantemente em atributos dimensionais, com destaque para a contagem de caracteres (`F2`: 719,27), sílabas (`F54`: 196,40) e palavras (`F1`: 100,07). Esse comportamento indica uma possível limitação conceitual: ao priorizar o comprimento do texto em detrimento de aspectos associados à intenção persuasiva, o modelo pode tornar-se mais suscetível a estratégias de evasão em cenários de produção, como a geração de mensagens de *phishing* mais curtas.

A fim de melhorar a capacidade de generalização do modelo, reduzindo o viés dimensional, densevolveu-se um novo corpus (*Dataset B*) com amostras menores, restritas a tamanhos de 80 a 100 palavras.

4.4. Desempenho sob Restrição Dimensional e a Relevância dos Verbos Imperativos

Para neutralizar o viés dimensional identificado na avaliação adversarial, procedeu-se ao treinamento e validação sobre o *Dataset* B. Como medida de segurança adicional de código, os quatro atributos diretamente atrelados à dimensão absoluta do texto — Contagem de Palavras (`F1`), Contagem de Caracteres (`F2`), Média de Letras por Palavra (`F3`) e Média de Palavras por Sentença (`F5`) — foram isolados e removidos da vetorização. Consequentemente, o modelo operou sobre um subconjunto estrito de 56 características. Forçado a buscar dependências puramente estilométricas sob este cenário restritivo, o classificador *XGBoost* manteve um nível de desempenho relevante, alcançando Acurácia de $88,0\%$ e *F1-Score* de $0,88$.

Essa abordagem viabilizou uma nova extração de importância (*Feature Importance*), revelando que a variável `F41` (Contagem de Verbos Imperativos) assumiu a maior relevância na superfície de decisão, concentrando cerca de $22,7\%$ de todo o peso preditivo, seguido pela frequência do caractere de barra (`F29`), com $8,90\%$, e pelo índice de legibilidade estrutural *SMOG* (`F32`), com $4,81\%$.

Esse fenômeno empírico corrobora as descobertas recentes de [Opara et al. 2025] sobre o peso preditivo do uso de verbos imperativos. Na literatura contemporânea de cibersegurança, demonstra-se que, em ataques de engenharia social, o ofensor frequentemente necessita persuadir a vítima a executar uma ação imediata (e.g., clicar em um *link* ou fornecer uma credencial), utilizando marcadores de comando para evocar um falso senso de urgência [Opara et al. 2025]. Ao rastrear, além dos imperativos explícitos, a morfologia dos verbos no modo subjuntivo atrelados à intenção de ação (conforme delineado na extração customizada da `F41`), o algoritmo demonstrou inicialmente capacidade de detectar a essência coercitiva da ameaça, independente da presença de gatilhos de *marketing* ou da extensão padronizada da mensagem.

4.5. Validação em Domínio Cruzado e Eficácia do Modelo Híbrido

A resiliência de um sistema de detecção de *phishing* em ambientes reais depende de sua capacidade de generalizar padrões entre diferentes arquiteturas geradoras. Para avaliar essa premissa, conduziu-se uma validação de domínio cruzado (*Cross-Domain Validation*). A capacidade preditiva do modelo *XGBoost*, treinado com o *Dataset B* (arquitetura *Gemini*), foi avaliada sobre o conjunto de dados inéditos, oriundos do estudo de [Opara et al. 2025] (arquitetura *GPT-4o*), os quais foram submetidos à mesma padronização e extração de características.

Os resultados desta validação externa indicaram uma degradação considerável na capacidade preditiva do classificador, com a Acurácia e o *F1-Score* macro decaindo para as margens de $0,52$ e $0,49$, respectivamente.

Uma investigação analítica revelou que o algoritmo sofreu os efeitos de um agudo *dataset shift* (desvio de dados). Para quantificar essa divergência, conduziu-se uma análise comparativa das médias de ativação dos atributos estilométricos nos e-mails maliciosos de ambos os *corpus*. Conforme detalhado na Tabela 4, as assinaturas sintáticas das duas IAs diferem significativamente.

Tabela 4. Análise de Desvio de Dados (*Dataset Shift*): Disparidade nas médias de ativação de características estilométricas em *phishings* gerados por *Gemini* e *GPT-4o*.

Atributo Estilométrico	Média *Gemini* (Treino)	Média *GPT-4o* (Teste)
`F58` (Marcadores de Urgência)	0,004	1,127
`F48` (Palavras Promocionais)	0,000	0,873
`F46` (Menções a Anexos)	0,012	0,698
`F40` (Proporção de Pronomes)	0,004	0,397

Os dados da Tabela 4 evidenciam o motivo do declínio preditivo. Os e-mails maliciosos de [Opara et al. 2025], gerados pelo *GPT-4o*, apresentaram uma incidência consideravelmente superior de marcadores de urgência (`F58`), palavras promocionais (`F48`) e

menções a anexos (F46). Como essas características possuíam peso ínfimo na árvore de decisão do modelo treinado no *Dataset B* (onde a arquitetura *Gemini* adotou uma postura mais sutil e sem jargões de *marketing*), o classificador não logrou êxito em reconhecer o ataque do *GPT-4o*.

Para solucionar essa limitação e desenvolver um sistema agnóstico ao modelo gerador, estruturou-se o *Dataset C (Híbrido)*, mesclando as instâncias de ambas as arquiteturas. Visando mitigar o superajuste (*overfitting*) sobre este novo *corpus* heterogêneo, o *XGBoost* foi submetido a um processo de *Fine-Tuning* utilizando busca em grade (*Grid Search*) combinada com Validação Cruzada Estratificada (*5-Fold Cross-Validation*).

A configuração otimizada estabilizou a decisão do classificador, que passou a apresentar $91{,}0\%$ de Acurácia e *F1-Score* de $0{,}91$ no conjunto de teste. A análise por origem dos dados indica leve variação de desempenho, com *F1-Score* de $0{,}92$ no subconjunto gerado pelo *Gemini* e $0{,}88$ no subconjunto do *GPT-4o*. A matriz de confusão (Tabela 5) registrou 2 falsos positivos e 1 falso negativo em um total de 26 amostras.

Tabela 5. Matriz de Confusão Isolada - Validação *Cross-Domain* (GPT-4o).

	Predito: Legítimo	**Predito: Phishing**
Real: Legítimo	11 (TN)	2 (FP)
Real: Phishing	1 (FN)	12 (TP)

A análise qualitativa dos erros observados, cujos trechos ilustrativos estão detalhados na Tabela 6, sugere padrões distintos de comportamento. O único falso negativo está associado a uma estratégia de evasão baseada em camuflagem linguística, na qual o modelo *GPT-4o* suprimiu a variável F41 (verbos imperativos), substituindo-a por construções persuasivas indiretas e de menor carga coercitiva.

Tabela 6. Trechos de amostras do *GPT-4o* classificadas incorretamente.

Erro do Modelo	**Trecho Sintético Extraído do Experimento**
Falso Negativo	*"Observe que essa oferta promocional será encerrada em breve. Pedimos que o senhor aja imediatamente para garantir que não perca esse desconto considerável."*
Falso Positivo 1	*"O senhor deve seguir estas etapas: 1. Acesse [...]. 2. Clique no botão de login [...]. 3. Digite seu e-mail e senha registrados."*
Falso Positivo 2	*"Recentemente, notamos uma atividade relacionada a uma transação financeira feita em sua conta. [...] Por favor, leia o documento em anexo o mais rápido possível."*

Por outro lado, os falsos positivos ocorreram em mensagens legítimas que apresentam características retóricas similares às observadas em ataques de *phishing*. Em um dos casos, tratava-se de uma comunicação de investimento que instruía explicitamente ações de *login*, resultando na forte ativação da variável F41. No segundo caso, a mensagem correspondia a um alerta financeiro solicitando a verificação de um arquivo anexado, o que ativou a variável F46 (menções a anexos).

Estes resultados reforçam que, em ambientes corporativos, comunicações legítimas frequentemente mimetizam a sintaxe de ataques. Portanto, a análise estilométrica atua com maior eficácia quando integrada como camada complementar a filtros tradicionais de metadados, e não como mecanismo de bloqueio isolado.

Notavelmente, a análise de importância de atributos (*Feature Importance*) do modelo híbrido revelou uma distribuição de pesos preditivos mais equilibrada para generalizar a ameaça entre diferentes IAs. O classificador consolidou a variável `F41` (Contagem de Verbos Imperativos e de Comando) como o preditor primordial, concentrando $15,88\%$ do Ganho de Informação. Os atributos de suporte mais relevantes passaram a ser a contagem de Marcadores de Urgência (`F58`) e o índice de legibilidade *SMOG* (`F32`), com $4,69\%$ e $4,62\%$, respectivamente.

Esse comportamento final sugere que, independentemente da arquitetura geradora, a estrutura base da engenharia social coercitiva sintética apoia-se fundamentalmente em diretivas de ação explícitas acopladas à urgência psicológica [Opara et al. 2025] e a modulações na complexidade de leitura do texto. Dessa forma, os resultados indicam que um sistema focado em estilometria comportamental apresenta potencial para atuar como uma camada de defesa adaptável e resiliente contra campanhas modernas de *phishing* geradas por Inteligência Artificial.

5. Conclusão

5.1. Principais Contribuições e Descobertas

Os resultados desta pesquisa oferecem indícios empíricos de que a estilometria comportamental é uma abordagem viável para complementar as defesas tradicionais. Contudo, os achados devem ser interpretados sob o escopo de uma prova de conceito em ambiente controlado. Portanto, o *framework* desenvolvido simula a eficácia da triagem de intenções maliciosas.

Apesar dessa limitação *in vitro*, as adaptações morfológicas desenvolvidas para o português brasileiro e os diagnósticos analíticos sobre o comportamento de algoritmos de detecção em cenários adversariais fornecem contribuições relevantes.

Em primeiro lugar, os resultados indicam que classificadores treinados com *prompts* pouco restritivos podem desenvolver uma aparente sensação de acurácia. Na ausência de palavras promocionais (`F48`), notou-se que o modelo tendeu a ancorar suas decisões em atributos dimensionais do texto (como a contagem de caracteres e de sílabas). A identificação dessa vulnerabilidade evidencia a importância de auditar os *datasets* sintéticos para mitigar o risco de que as defesas baseadas em IA sejam evadidas por alterações simples na extensão das mensagens.

Além disso, ao isolar o viés dimensional e o vocabulário de *marketing* (por meio da restrição de 80 a 100 palavras aplicada ao *Dataset B*), o algoritmo *XGBoost* passou a priorizar características morfológicas e estruturais. A variável `F41` (Contagem de Verbos Imperativos) emergiu como o atributo de maior peso na superfície de decisão, concentrando aproximadamente $22,7\%$ do ganho de informação. Esse comportamento empírico corrobora as discussões recentes da literatura [Opara et al. 2025], sugerindo que a engenharia social coercitiva está frequentemente associada ao uso de verbos de ação que buscam persuadir o usuário.

Por fim, a pesquisa mapeou evidências de um desvio de domínio (*dataset shift*) ao avaliar o classificador treinado sobre dados do *Gemini* contra ameaças geradas pelo *GPT-4o*. Observou-se que as arquiteturas geradoras avaliadas apresentaram assinaturas estilométricas distintas, com destaque para a maior incidência de marcadores de urgência

na IA da OpenAI. A proposição de um modelo híbrido (*Dataset C*) apresentou resultados relevantes na mitigação dessa limitação, alcançando um *F1-Score* global de $0,91$. Para contornar as disparidades estruturais entre as IAs, o modelo evitou a dependência excessiva de uma única variável e pulverizou o peso preditivo, consolidando a contagem de verbos imperativos (`F41`) como a principal assinatura comportamental observada, atuando em conjunto com índices de complexidade de leitura (`F32`) e gatilhos de urgência (`F58`).

5.2. Limitações e Trabalhos Futuros

Apesar da eficácia na mitigação parcial do *dataset shift*, a diversidade de arquiteturas avaliadas restringiu-se aos modelos *Gemini* e GPT-4o. Pesquisas futuras podem expandir o *corpus* para *LLMs* abertos (ex: *LLaMA* e *Claude*) para validar a consistência das assinaturas estilométricas contra ameaças *zero-day*. Ademais, a variabilidade dimensional e linguística limitou-se a textos curtos (80 a 100 palavras) calibrados estritamente para o Português do Brasil. Cenários reais exigem testar e-mails extensos ou composições híbridas (humano-IA), além de reconfigurar o *pipeline* morfológico para outros idiomas. Portanto, trabalhos futuros exigem a validação com dados orgânicos (reais) integrados a *gateways* corporativos em tempo real. Além de atestar a latência computacional e escalabilidade, esse teste real é necessário para mensurar falsos positivos, visto que comunicações legítimas podem mimetizar a sintaxe imperativa de ataques.

Referências

Alhaji, U. M., Adewumi, S. E., and Yemi-peters, V. I. (2025). Classification of phishing attacks using machine learning algorithms: A systematic literature review. *Journal of Advances in Mathematics and Computer Science*, 40(1):26–44.

Barik, K., Misra, S., et al. (2025). Web-based phishing url detection model using deep learning optimization techniques. *International Journal of Data Science and Analytics*, 20:4449–4471.

Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., and Gollangi, H. K. (2024). AI-driven phishing email detection: Leveraging big data analytics for enhanced cybersecurity. *Library Progress International*, 44(3):7211–7224.

Eze, C. S. and Shamir, L. (2024). Analysis and prevention of AI-based phishing email attacks. *Electronics*, 13(10):1839.

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

Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. G. V., and Jr., O. N. O. (1996). Readability formulas applied to textbooks in brazilian portuguese. Technical Report 28, Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, SP, Brazil.

Opara, C., Modesti, P., and Golightly, L. (2025). Evaluating spam filters and stylometric detection of AI-generated phishing emails. *Expert Systems With Applications*, 276:127044.

Owa, K. and Adewole, O. (2025). Benchmarking machine learning techniques for phishing detection and secure URL classification. *International Journal of Computer Science and Mobile Computing*, 14(1):20–37.

Patil, R. R. et al. (2025). Design of intelligent feature selection technique for phishing detection. *IIUM Engineering Journal*, 26(1).

Patil, R. R., Kaur, G., Jain, H., Tiwari, A., Joshi, S., Rao, K., and Sharma, A. (2022). Machine learning approach for phishing website detection: A literature survey. *Journal of Discrete Mathematical Sciences and Cryptography*.

Saeed, E. M. H. (2025). An ensemble voting classifier based on machine learning models for phishing detection. *International Journal of Scientific Research in Science, Engineering and Technology*, 12(1):15–27.

Silva, R. R. and Pires, T. B. (2024). Olhares sobre a tradução automática: uma análise sobre o desempenho dos sistemas deepl e google tradutor. *Revista Babel*, 14:e20368.

Tang, L. and Mahmoud, Q. H. (2021). A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction*, 3:679–702.

A. Apêndice: Dicionário de Atributos Estilométricos

A Tabela 7 detalha a composição de todas as 60 características extraídas durante a etapa de Processamento de Linguagem Natural, organizadas em oito dimensões analíticas.

Tabela 7. Descrição e categorização dos 60 atributos estilométricos extraídos (F1 a F60).

Categoria (IDs)	Descrição dos Atributos
Lexicais (F1–F12)	Total de palavras e caracteres, média de letras por palavra, contagem de sentenças, média de palavras por sentença, diversidade lexical (palavras únicas), incidência de símbolos '@', frequência e proporção de palavras em letras maiúsculas, palavras complexas (> 6 letras) e média silábica.
Sintáticos e POS (F13–F20, F36–F40)	Contagem intra-frasal de pontuações (vírgula, ponto e vírgula, dois-pontos), densidade de conjunções, densidade de orações subordinadas (`F17`), densidade de pronomes, preposições, palavras de função e distribuição proporcional das classes gramaticais (substantivos, verbos, adjetivos e advérbios).
Pontuação (F21–F30)	Frequência normalizada de 9 marcadores pontuais e caracteres especiais comuns (., !, ?, -, ", /, etc.) e um índice global de variedade de pontuação.
Legibilidade (F31–F35)	Índices padronizados de esforço cognitivo de leitura: *Flesch Reading Ease* (`F31` - ajustado), *SMOG Index*, *Dale-Chall*, *Coleman-Liau* e *Gunning Fog*.
Intenção e Comando (F41–F45)	Frequência de verbos de ação focados em comando (`F41`), verbos modais (ex: "deve", "pode"), advérbios de incerteza (ex: "talvez") e a contagem de pronomes focados na 1ª, 2ª e 3ª pessoa.
Jargões e Metadados (F46–F50)	Menções textuais a arquivos anexos/faturas, incidência de jargões técnicos (ex: "autenticação", "senha"), palavras promocionais (`F48`), e a extensão absoluta (palavras e caracteres) do Assunto do e-mail.
Complexidade (F51–F55)	Contagem de combinações n-gramas frequentes (bigramas e trigramas), desvio padrão da variação de tamanho das palavras, contagem total de sílabas e polissílabas.
Estilísticos (F56–F60)	Incidência de marcadores de polidez, agressividade coercitiva, indicadores de extrema urgência (`F58`), orações condicionais (ex: "caso", "se") e termos de personalização direta (ex: "prezado", "você").

B. Apêndice: Engenharia de Prompts para Geração dos Datasets

A Tabela 8 detalha as instruções literais fornecidas à interface da API dos Modelos de Linguagem para a síntese das mensagens. As variáveis entre colchetes, como `[tema]` e `[ação]`, foram preenchidas dinamicamente nos scripts.

Tabela 8. Diretivas de *prompt* utilizadas para a geração dos *datasets*.

Categoria e Nível	Instrução Literal Fornecida ao LLM
	***Dataset* A: Abordagem Orgânica (Baseline) [Opara et al. 2025]**
Phishing (Básico)	"Gere um e-mail de phishing simples pedindo para o usuário `[ação]` sobre o tema `[tema]`."
Phishing (Interm.)	"Crie um e-mail de phishing com formatação adequada e um logotipo de empresa, solicitando que o usuário `[ação]` sobre o tema `[tema]`."
Phishing (Avanç.)	"Gere um e-mail de phishing altamente sofisticado com linguagem urgente, um endereço de remetente forjado e formatação profissional pedindo para o usuário `[ação]` sobre o tema `[tema]`."
Legítimo (Corp.)	"Gere um e-mail profissional de uma organização respeitável informando o destinatário sobre uma oferta promocional ou acordo em `[tema]` que ele possa considerar valioso. Incentive-o a saber mais através de clicar no link. O e-mail deve ser claro, amigável e confiável."
	***Dataset* B: Abordagem Restritiva (*Red Team*)**
Phishing (Evasivo)	"Gere um e-mail de phishing sobre `[tema]`. Escreva de forma funcional e direta, simulando tratativa corporativa. IMPORTANTE: 1. O texto deve ter obrigatoriamente entre 80 e 100 palavras. 2. Não utilize termos promocionais ou de urgência extrema (ex: 'ganhe', 'oferta', 'promoção', 'urgente'). 3. Não inclua metadados como 'Assunto:', 'Corpo:', ou termos como 'Link de phishing', 'Link malicioso' e/ou 'link falso'. 4. Retorne apenas o texto limpo do e-mail, sem feedbacks ou comentários da IA."
Legítimo (Técnico)	"Gere um e-mail profissional legítimo sobre `[tema]`. O texto deve ser técnico, formal e seguir o padrão de uma organização respeitável. IMPORTANTE: 1. O texto deve ter obrigatoriamente entre 80 e 100 palavras. 2. Não utilize linguagem de marketing, vendas ou incentivos promocionais. 3. O email dever ser claro, amigável e confiável. 4. Retorne apenas o texto limpo do e-mail, sem feedbacks, comentários ou o campo Assunto."