GUARDIAN AI

Guia de Análise por E-mail

Sistema Guardian Forward

Como encaminhar mensagens suspeitas, entender a análise e interpretar o relatório de risco.

Versão 1.0 | Julho de 2026

Endereço de análise: check@fraudguardian.com.br

fraudguardian.com.br

0. Comece aqui

Uma visão rápida para usar o Guardian sem precisar conhecer detalhes técnicos.

> **Uso em quatro passos**
>
> Abra a mensagem suspeita, escolha Encaminhar, envie para check@fraudguardian.com.br e leia o relatório retornado pelo Guardian. Não clique nos links nem abra anexos para realizar a análise.

Passo a passo

1. Abra o e-mail que você considera suspeito.
2. Use a função Encaminhar do seu provedor de e-mail.
3. Envie a mensagem para check@fraudguardian.com.br.
4. Aguarde o processamento automático e a resposta com o relatório.
5. Leia primeiro o veredito, depois a nota de risco, a confiança e as principais evidências.

Leitura em 20 segundos

0-10	35-96%	3
NOTA DE RISCO	CONFIANÇA ATUAL	FAIXAS DE RISCO

Nota de risco responde: “quão preocupante é a combinação de sinais encontrados?” Confiança responde: “quanto material consistente o Guardian conseguiu reunir para sustentar essa avaliação?”

> **Regra de segurança**
>
> Mesmo quando o relatório indicar baixo risco, não forneça senhas, códigos de autenticação, dados bancários ou informações pessoais sem confirmar a solicitação em um canal oficial.

Sobre o encaminhamento

O Guardian consegue analisar o conteúdo encaminhado, links e anexos. Entretanto, provedores de e-mail podem alterar ou remover parte dos cabeçalhos da mensagem original durante um encaminhamento comum. Quando isso acontece, algumas verificações de identidade podem ter menos informação disponível.

Sumário

Estrutura desta documentação.

> **Escopo desta versão**
>
> Este documento descreve o comportamento da implementação atual do Guardian AI em julho de 2026. Como o projeto está em evolução, módulos e critérios podem mudar entre versões.

1. O que é o Guardian Forward

O canal de análise por e-mail do Guardian AI.

Guardian Forward é o fluxo no qual uma pessoa encaminha uma mensagem suspeita para o Guardian e recebe um relatório automático. O objetivo é reduzir a necessidade de o usuário interpretar sozinho cabeçalhos, domínios, links e anexos.

O sistema não trabalha com uma única regra. Ele combina várias evidências independentes: linguagem do texto, reputação, autenticação de e-mail, estrutura de links, redirecionamentos, imitação de marcas, modelos de IA e análise de arquivos.

O que o Guardian pretende responder

- A mensagem apresenta sinais comuns de phishing ou engenharia social?
- O remetente e os domínios parecem coerentes com a organização citada?
- Os links levam para destinos diferentes ou visualmente semelhantes a domínios oficiais?
- Os anexos possuem sinais de risco, reputação conhecida ou detecções de malware?
- As evidências, quando combinadas, justificam uma classificação de risco baixa, média ou alta?

> **Importante**
>
> O Guardian produz uma estimativa de risco e uma explicação. Ele não certifica que uma mensagem é legítima e não substitui confirmação com a instituição envolvida.

2. Como o sistema funciona

Da caixa de entrada ao relatório devolvido ao usuário.

1	2	3	4	5
Recebimento	**Extração**	**Análise**	**Score**	**Relatório**
IMAP recebe a mensagem	Texto, links e anexos	Regras, IA e reputação	Risco e confiança	HTML + texto e resposta

2.1 Recebimento por IMAP

A caixa de análise é monitorada por IMAP. Quando uma nova mensagem elegível é encontrada, o Guardian obtém os dados necessários para o processamento e evita reprocessar mensagens já tratadas.

2.2 Extração e armazenamento

O coletor identifica remetente, assunto, corpo da mensagem, links, PDFs, imagens e demais arquivos. Esses elementos são associados à análise para que cada módulo possa trabalhar com a evidência adequada.

2.3 Motor de análise

O guardian_analyzer reúne as evidências produzidas pelos diferentes módulos. A ausência de um módulo opcional, como um modelo BERTimbau ainda não treinado, não interrompe o fluxo: o restante da análise continua funcionando.

2.4 Relatório e resposta

Após calcular o resultado, o Guardian cria um relatório com score, confiança, evidências, recomendações e um código público. O renderer produz uma versão HTML e uma versão em texto puro para compatibilidade com clientes de e-mail, e a resposta é enviada via SMTP.

3. O que é analisado

As principais camadas de evidência do sistema.

Conteúdo textual	Urgência, ameaça, solicitação de credenciais, pagamento, recompensa, pressão para clicar ou agir.
Remetente e reputação	Domínio, listas de reputação e coerência entre a organização citada e a origem da mensagem.
Cabeçalhos	SPF, DKIM, DMARC, Reply-To, Return-Path, Message-ID e cadeia Received quando disponíveis.
Links	Estrutura da URL, domínio, reputação, redirecionamentos e destino final.
Typosquatting	Semelhança com domínios oficiais e sinais de imitação de marcas conhecidas.
PDFs e imagens	Hash, reputação e, no caso de PDF, análise textual pelo modelo específico quando disponível.
Arquivos	Hash SHA-256, metadados e consulta de malware/reputação.
BERTimbau	Classificação textual como evidência adicional, não como decisão isolada.
Correlação	E-mails semelhantes podem ser relacionados a campanhas quando atingem os critérios internos.

Nenhuma dessas características deve ser interpretada como prova isolada de fraude. O valor do Guardian está na correlação das evidências.

4. Como o score é calculado

Risco, evidências protetivas e confiança são conceitos diferentes.

4.1 Nota de risco

Cada evidência possui um valor de risco e uma confiança própria. A contribuição ponderada é calculada como risco x confiança. Evidências de proteção têm valor negativo e podem reduzir a soma. O resultado final é limitado ao intervalo de 0 a 10.

> **Fórmula simplificada**
> Score de risco = soma das evidências de risco ponderadas - soma das evidências protetivas ponderadas, limitado entre 0 e 10.

Nota de risco	Classificação	Interpretação prática
0,0 a 3,49	**Pouca chance de golpe**	Poucos sinais fortes foram encontrados. Mantenha cautela.
3,5 a 6,99	**Chance média de golpe**	Há sinais de alerta. Confirme a mensagem em um canal oficial antes de agir.
7,0 a 10,0	**Grande chance de golpe**	Há uma combinação forte de evidências. Não interaja antes de verificar a origem.

4.2 Confiança

A confiança não é a probabilidade matemática de o e-mail ser um golpe. Na implementação atual, ela cresce com a quantidade e a força das evidências disponíveis, começa em 35% e é limitada a 96%.

> **Exemplo**
> Risco 8,0/10 com 55% de confiança indica um resultado preocupante, mas sustentado por menos evidências do que risco 8,0/10 com 92% de confiança. Em ambos os casos, a atitude segura é verificar a origem antes de interagir.

5. Como interpretar o relatório

A anatomia do documento recebido por e-mail.

01	**Veredito**	Síntese visual do resultado geral.
02	**Nota de risco**	Valor de 0 a 10 produzido pela combinação das evidências.
03	**Confiança**	Quanto material consistente o Guardian conseguiu reunir.
04	**Dados analisados**	Remetente, assunto e quantidade de links, PDFs, imagens e arquivos.
05	**Principais evidências**	Sinais de risco e evidências protetivas que contribuíram para o resultado.
06	**Links e PDFs**	Detalhes dos elementos extraídos e avaliados.
07	**Malware Analysis**	Resultado de reputação e antivírus quando há arquivos analisáveis.
08	**Recomendações**	Ações sugeridas para reduzir o risco.
09	**Análise completa**	Link para o relatório público detalhado quando disponível.

5.1 Veredito

O veredito traduz o score para linguagem direta: Pouca chance de golpe, Chance média de golpe ou Grande chance de golpe. Quando o resultado não puder ser determinado adequadamente, a interface também possui um estado inconclusivo.

5.2 Principais evidências

As evidências são ordenadas para destacar os sinais com maior impacto. Elas podem indicar risco, como “Reply-To diferente do remetente”, ou proteção, como uma autenticação válida. Leia o conjunto e não apenas uma linha isolada.

5.3 Dados analisados

Essa seção ajuda a confirmar o escopo: remetente, assunto e quantidade de links, PDFs, imagens e arquivos identificados. Se algo esperado não aparecer, a análise pode ter recebido uma versão incompleta da mensagem.

5.4 Recomendações

As recomendações são adaptadas ao contexto. Mensagens com links recebem orientação para verificar o domínio; mensagens com PDFs recebem atenção adicional a boletos e anexos; risco alto prioriza a instrução de não interagir antes de confirmar com a instituição.

5.5 Relatório completo

Quando o relatório possui URL pública, o botão de análise completa leva a uma página identificada por um código aleatório. Essa página pode apresentar detalhes adicionais do score e dos módulos utilizados.

6. Identidade do e-mail

SPF, DKIM, DMARC e inconsistências de cabeçalhos.

O Guardian pode usar os cabeçalhos armazenados para procurar sinais de falsificação e inconsistência. Esses mecanismos ajudam, mas não funcionam como certificado absoluto de legitimidade.

SPF: Verifica se o servidor de envio estava autorizado, conforme a política publicada pelo domínio.

DKIM: Verifica uma assinatura criptográfica adicionada ao e-mail.

DMARC: Combina alinhamento de SPF/DKIM com a política definida pelo proprietário do domínio.

Reply-To: Indica para onde a resposta será enviada. Uma diferença inesperada em relação ao From pode ser relevante.

Return-Path: Endereço técnico usado para retornos de entrega; diferenças podem ser legítimas, mas entram como sinal contextual.

Message-ID: Identificador da mensagem. Um domínio diferente recebe peso baixo porque provedores terceirizados podem gerar esse campo.

> **Cuidado com encaminhamentos**
>
> Ao encaminhar uma mensagem, seu provedor pode substituir os cabeçalhos originais pelos cabeçalhos do encaminhamento. Nesse cenário, o Guardian pode ter menos dados para avaliar a identidade do remetente original.

7. Links, redirecionamentos e typosquatting

Como o Guardian procura destinos escondidos e imitações de marcas.

7.1 Análise da URL

O sistema normaliza os links encontrados e avalia o domínio, a reputação e características estruturais. O objetivo é descobrir sinais que não são óbvios para quem vê apenas o texto do e-mail.

7.2 Cadeia de redirecionamento

O analisador pode seguir redirecionamentos HTTP/HTTPS de forma controlada, com limite de saltos e bloqueio de destinos de rede privada para reduzir risco de SSRF. Múltiplos redirecionamentos ou mudança do domínio inicial para outro domínio adicionam evidências de risco.

> **Exemplo de leitura**
>
> Um link pode começar em dominio-a.com, passar por um encurtador e terminar em dominio-b.xyz. O Guardian registra a cadeia e considera a diferença entre origem e destino final.

7.3 Typosquatting

O módulo de typosquatting compara o domínio observado com domínios oficiais cadastrados. Ele considera semelhança textual, presença do nome da marca dentro de um domínio não oficial, Punycode e uso excessivo de hífens.

Exemplo didático: um domínio visualmente parecido com nubank.com.br, mas registrado em outro endereço, pode ser sinalizado como possível imitação. A conclusão final ainda depende das demais evidências.

8. Arquivos, PDFs e Malware Analysis

O que acontece quando a mensagem contém anexos.

Arquivos associados ao e-mail podem receber um hash SHA-256, que funciona como uma impressão digital do conteúdo. Esse identificador permite verificar reputação local e consultar serviços externos sem depender apenas do nome do arquivo.

8.1 VirusTotal

Quando a integração está habilitada, o Guardian consulta o VirusTotal pela hash do arquivo. O relatório pode mostrar quantos mecanismos classificaram o arquivo como malicioso, suspeito, inofensivo ou não detectado, além de detalhes de mecanismos individuais.

> **Privacidade**
>
> A implementação do Guardian permite trabalhar somente com a consulta de hash. O envio de arquivos desconhecidos ao VirusTotal é uma opção de configuração e deve ser tratado de acordo com a política de privacidade do ambiente.

8.2 Como interpretar “nenhum malware detectado”

Esse resultado significa apenas que, naquele momento e naquela consulta, os mecanismos disponíveis não sinalizaram o arquivo. Malware novo, ofuscado ou direcionado pode não ser reconhecido. Portanto, ausência de detecção não é garantia de segurança.

8.3 PDFs

Além de reputação e malware, PDFs podem fornecer texto para análise. Quando um modelo pdf_bert está ativo, o conteúdo textual extraído pode ser classificado como uma evidência adicional.

9. Inteligência Artificial e BERTimbau

Como os modelos treinados entram na decisão.

O Guardian suporta duas famílias de modelos BERTimbau: email_bert, voltado a textos de e-mail, e pdf_bert, voltado ao texto extraído de PDFs. Cada família possui versões e um arquivo current.json indicando o modelo ativo.

Durante a análise, o modelo só é utilizado quando está habilitado e disponível. Textos muito curtos podem ser ignorados, e a saída do modelo entra no conjunto de evidências em vez de substituir as regras, a reputação ou os demais módulos.

> **Por que isso é importante?**
>
> Uma frase aparentemente suspeita pode existir em um e-mail legítimo. Da mesma forma, um golpe sofisticado pode usar linguagem neutra. Combinar IA com sinais técnicos reduz a dependência de uma única fonte de decisão.

O que fazer quando não há modelo ativo

Nada especial é exigido do usuário. A análise continua com as regras explicáveis, reputação, links, cabeçalhos e arquivos. O relatório não deve ser interpretado como incompleto apenas porque um modelo opcional não participou; a confiança e as evidências disponíveis ajudam a contextualizar o resultado.

10. O que fazer depois do resultado

Ações práticas para cada faixa de risco.

POUCA CHANCE DE GOLPE

- Mantenha atenção a links e anexos inesperados.
- Se houver pedido de senha, pagamento ou código, confirme mesmo assim em um canal oficial.
- Verifique se o domínio e a situação fazem sentido para você.

CHANCE MÉDIA DE GOLPE

- Não tome decisões financeiras ou de acesso imediatamente.
- Abra o site oficial digitando o endereço manualmente, em vez de usar o link recebido.
- Entre em contato com a instituição por telefone ou aplicativo oficial.

GRANDE CHANCE DE GOLPE

- Não clique em links, não abra anexos e não responda até confirmar a origem.
- Não forneça senhas, códigos, documentos ou dados bancários.
- Considere bloquear o remetente e denunciar a tentativa ao provedor ou à instituição imitada.

> **Situação crítica**
> Se você já informou senha, código de autenticação, dados bancários ou realizou pagamento após interagir com uma mensagem suspeita, a prioridade deixa de ser apenas analisar o e-mail. Altere credenciais, contate a instituição envolvida pelos canais oficiais e siga os procedimentos de segurança aplicáveis.

11. Limitações, privacidade e boas práticas

Como usar o resultado de forma responsável.

11.1 Limitações

- Modelos de IA podem errar e a linguagem de golpes muda com frequência.
- A qualidade da análise depende dos dados disponíveis na mensagem recebida.
- Links podem apresentar conteúdo diferente por horário, região, dispositivo ou sessão.
- Arquivos maliciosos novos podem não ter reputação ou detecção conhecida.
- Mensagens legítimas podem gerar falsos positivos quando possuem sinais parecidos com golpes.
- Uma mensagem sofisticada pode produzir falso negativo se não apresentar evidências suficientes.

11.2 Dados sensíveis

Evite adicionar ao encaminhamento qualquer senha, código de autenticação, número completo de cartão, documento pessoal ou dado altamente sensível que não seja necessário para a análise. O conteúdo original da mensagem será processado conforme a configuração do Guardian.

11.3 Serviços externos

Algumas funções, como VirusTotal, podem depender de serviços de terceiros. A configuração do ambiente define se apenas hashes são consultados ou se arquivos desconhecidos podem ser submetidos para análise externa.

> **Princípio de uso**
> O relatório deve apoiar a decisão, nunca substituir uma verificação oficial quando houver dinheiro, credenciais, identidade ou acesso a contas em jogo.

12. Perguntas frequentes

Respostas rápidas para situações comuns.

O Guardian disse “baixo risco”. Posso clicar?

Baixo risco significa que o sistema não encontrou sinais fortes suficientes. Não é garantia de legitimidade. Se a mensagem pedir dados sensíveis ou dinheiro, confirme por outro canal.

O que significa confiança alta?

Significa que o Guardian reuniu evidências mais fortes e/ou numerosas para sustentar a análise. Não significa certeza absoluta.

Por que um e-mail legítimo pode receber risco médio?

Cobranças, urgência, links externos e diferenças de infraestrutura também aparecem em comunicações reais. O Guardian prefere mostrar os sinais para que o usuário verifique o contexto.

Por que não apareceu Malware Analysis?

A seção pode não aparecer quando não há arquivos analisáveis, quando a integração externa está desabilitada ou quando não houve resultado disponível.

SPF, DKIM e DMARC válidos tornam o e-mail seguro?

Não. Eles ajudam a validar a origem técnica, mas uma conta legítima comprometida ou um domínio malicioso bem configurado ainda pode enviar mensagens perigosas.

O BERTimbau decide sozinho?

Não. O modelo é uma evidência entre várias. A decisão final combina IA com regras, reputação e sinais técnicos.

Posso encaminhar qualquer mensagem?

Tecnicamente, o sistema foi pensado para analisar mensagens suspeitas. Evite encaminhar conteúdo desnecessariamente sensível ou dados de terceiros sem necessidade.

O relatório pode mudar no futuro?

Sim. Reputação, detecções antivírus, conteúdo de sites e os próprios modelos do Guardian podem mudar com o tempo.

13. Glossário

Termos técnicos usados nesta documentação e nos relatórios.

Termo	Significado
Phishing	Tentativa de enganar uma pessoa para que entregue informações, dinheiro ou acesso.
Typosquatting	Registro de domínio parecido com o oficial para induzir erro visual.
SPF	Mecanismo que ajuda a verificar se o servidor estava autorizado a enviar e-mail pelo domínio.
DKIM	Assinatura criptográfica usada para verificar integridade e autenticidade da mensagem.
DMARC	Política que usa SPF e DKIM para orientar a validação de mensagens de um domínio.
SHA-256	Hash criptográfico usado como identificador do conteúdo de um arquivo.
Redirect	Redirecionamento de uma URL para outro endereço.
BERTimbau	Modelo de linguagem em português usado pelo Guardian como uma das evidências de análise textual.
VirusTotal	Serviço externo que agrega resultados de diferentes mecanismos de segurança para arquivos.
Falso positivo	Conteúdo legítimo classificado como suspeito.
Falso negativo	Conteúdo malicioso que não foi identificado como de alto risco.

GUARDIAN AI

Guardian Forward

Análise de e-mails suspeitos baseada em evidências, explicabilidade e educação digital.

check@fraudguardian.com.br

fraudguardian.com.br

Versão 1.0 | Julho de 2026

Esta documentação descreve a implementação atual do Guardian AI e pode ser atualizada conforme o projeto evolui.