

Dicas para detecção de IA

Com o rápido desenvolvimento das tecnologias de IA generativa, as ferramentas de detecção de IA continuam sendo um recurso importante para verificar a autenticidade do conteúdo audiovisual e combater as ameaças representadas pelo uso enganoso de mídia sintética. [Este guia reúne dicas e práticas úteis compartilhadas por verificadores de fatos e jornalistas sobre como usar ferramentas de detecção de IA de maneira eficaz.](#)

Este documento foi elaborado como parte dos esforços mais amplos da WITNESS para promover a detecção eficaz e inovadora de IA. Ele acompanha [TRIED: Um benchmark de detecção por IA verdadeiramente inovador e eficaz](#) (disponível em inglês) que redefine a eficácia das ferramentas de detecção de IA com base em seu impacto no mundo real.

Por meio de conversas com parceiros apoiados pela equipe Força de Resposta Rápida a Deepfakes da WITNESS, foram coletados e compilados uma lista de recomendações práticas para fortalecer ainda mais as comunidades que lidam com casos de conteúdo suspeito de ter sido gerado por IA, com seis considerações-chave:

DICA #1: Desempenho em condições reais

1

Rastreie a versão mais antiga do conteúdo.

Utilize a pesquisa reversa de imagens e prints dos principais frames e imagens de vídeos para identificar conteúdos que possam ter sido colocados fora de contexto, detectar retocações ou pequenas edições, ou encontrar a versão de melhor qualidade. Apesar do [Google Lens](#) ser uma boa ferramenta, utilizar outras como o [Yandex](#) ou o [TinEye](#) pode ajudar a comparar os resultados para maior eficácia.

2

Leve em consideração os falsos positivos e os falsos negativos.

Lembre-se de que as ferramentas de detecção de IA não conseguem diferenciar entre edições fraudulentas e não fraudulentas. Gráficos sobrepostos e efeitos de edição, como o desfoque, podem gerar falsos positivos, mesmo quando não são adicionados com intenção maliciosa. Da mesma forma, um resultado negativo, por si só, não garante a autenticidade e deve ser avaliado em conjunto com outras evidências contextuais.

3

Verifique os requisitos da ferramenta.

Cheque se a ferramenta estabelece um comprimento mínimo para os arquivos de áudio ou vídeo que ela pode analisar, pois alguns cliques podem ser curtos demais para fornecer um resultado confiável.

DICA #2: Transparência e explicabilidade

1

Trate os resultados como sinais, não como verdades absolutas.

Leia os resultados fornecidos pela ferramenta com cautela.

2

Utilize uma lista de verificação pré-análise.

Antes de analisar os resultados, pergunte-se: o que você sabe sobre a ferramenta e os dados com os quais ela foi treinada? Que tipos de mídia e manipulações ela afirma ser capaz de detectar? Qual é o seu escopo e quais são suas possíveis limitações? Que tipo de evidência ela fornece (índices de confiança, artefatos, localização, detalhes do modelo)?

3

Criar protocolos de comunicação.

Estabelecer protocolos internos para a comunicação de resultados em diferentes tipos de casos, como, por exemplo, o que fazer caso diferentes ferramentas apresentem resultados conflitantes.

DICA #3: Acessibilidade e usabilidade direcionadas

1

Compare versões específicas para cada plataforma.

Baixe o conteúdo de diferentes plataformas (por exemplo, TikTok, X, Instagram, YouTube) e envie cada versão para o detector a fim de comparar os resultados (algumas ferramentas de detecção podem ter sido treinadas para o formato de arquivo do YouTube, mas não para o formato do TikTok).

2

Utilizar amostras de controle.

Antes de testar um item suspeito, envie um conteúdo cuja autenticidade seja comprovada e que tenha características semelhantes (tipo de arquivo, resolução, iluminação, tema) para avaliar o comportamento de referência da ferramenta.

3

Realizar validação cruzada com várias ferramentas.

Utilize várias ferramentas de detecção para validar suas descobertas de forma cruzada e comparar resultados.

4

Compartilhe o que você aprendeu.

Considere criar uma base de conhecimento para reunir e compartilhar as percepções obtidas por meio de seus testes. Isso ajudará a capacitar sua comunidade a lidar com diferentes ferramentas de detecção de IA de maneira informada e confiante.

DICA #4: Equidade e representatividade

1

Planeje-se para um desempenho variável em diferentes contextos.

Estabeleça protocolos internos para lidar com diferentes tipos de casos, incluindo arquivos de mídia diversos, áudios em diferentes idiomas, conteúdos voltados para diferentes públicos-alvo e provenientes de diferentes contextos regionais, com base na sua experiência no uso de diferentes ferramentas.

2

Pratique o jornalismo responsável.

Caso os resultados fornecidos pelas ferramentas possam ter sido afetados por limitações relacionadas à equidade, mencione-as em seu relatório para apresentar um relato transparente e abrangente do seu processo de verificação.

DICA #5: Durabilidade e resiliência

1

Verifique a manutenção e a atualidade das ferramentas.

Verifique quando a ferramenta foi atualizada pela última vez para garantir que ela seja capaz de acompanhar os avanços recentes em IA generativa.

2

Teste novamente as ferramentas que você deixou de priorizar.

Se você decidiu não usar determinadas ferramentas devido a resultados imprecisos consecutivos no passado, considere reavaliá-las para verificar se foram atualizadas desde então. Se sim, considere testá-las novamente para ver se as atualizações melhoraram o desempenho.

DICA #6: Integração em ecossistemas de verificação mais amplos

1

A detecção é uma etapa da verificação.

As ferramentas de detecção por IA não podem fornecer uma resposta completa sobre a autenticidade de um conteúdo e constituem apenas uma etapa no fluxo de trabalho de verificação. Confie, em primeiro lugar, em outros métodos de verificação e na sua experiência, e utilize as ferramentas para complementar o seu processo.

2

Recorra à experiência de especialistas, sempre que possível.

Quando os riscos forem elevados ou os sinais forem contraditórios, analise os resultados com um especialista capaz de interpretar os dados e ajudar a identificar possíveis falsos positivos ou falsos negativos.

Agradecimentos

WITNESS

Este guia de dicas baseia-se diretamente nas percepções coletadas durante o desenvolvimento do **TRIED: Um benchmark de detecção por IA verdadeiramente inovador e eficaz (versão disponível em inglês)**. Gostaríamos de agradecer aos jornalistas, verificadores de fatos e pesquisadores que generosamente compartilharam seu tempo e conhecimento: Rabi Alhassan (FactSpace West Africa), Tracy Aminu (FactSpace West Africa), Ana Bregvadze (GeoFacts) e Maiko Ratiani (Myth Detector). Também gostaríamos de agradecer aos parceiros que contribuíram com comentários, mas preferiram permanecer anônimos.

Encontre mais recursos



Encontre mais recursos

Este guia prático oferece informações importantes sobre o uso de ferramentas de detecção baseadas em IA para imagens, vídeos, áudio e texto, incluindo suas capacidades e limitações.



Visão geral do TRIED: o benchmark de detecção por IA verdadeiramente inovador e eficaz da WITNESS

Com base em experiências na linha de frente, casos de IA enganosa e consultas globais, o relatório descreve como as ferramentas de detecção devem evoluir para se tornarem verdadeiramente inovadoras e relevantes, atendendo a diversos contextos linguísticos, culturais e tecnológicos. (Disponível em inglês)



Biblioteca Online WITNESS

Acesse todos os recursos sobre verificação de mídia criados pela WITNESS.

Com o apoio de



Digital Democracy
Initiative