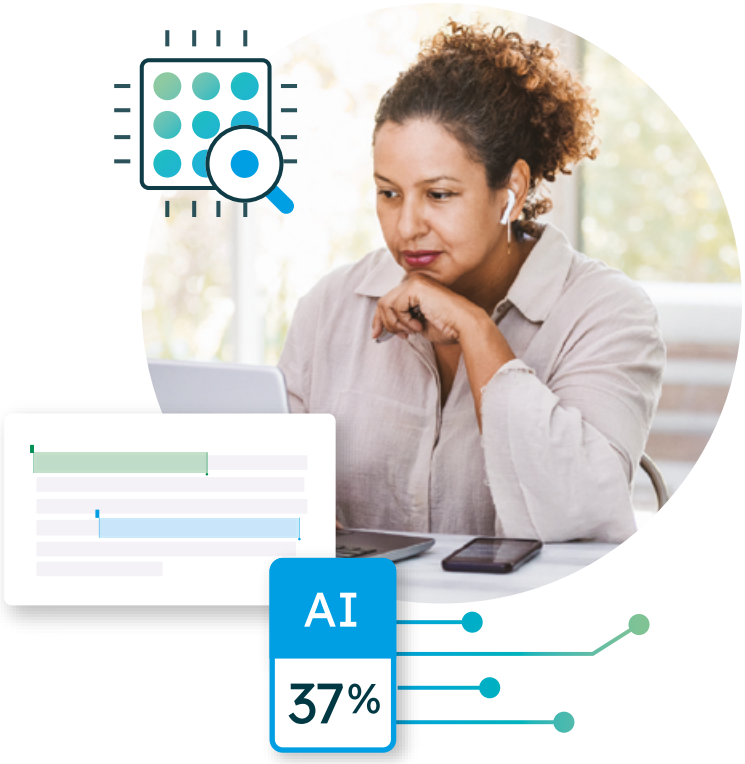
Perguntas frequentes sobre os recursos de detecção de escrita por IA da Turnitin



Como funcionam os recursos de detecção de escrita por IA da Turnitin?

1. A Turnitin oferece uma solução para detectar escrita por IA?

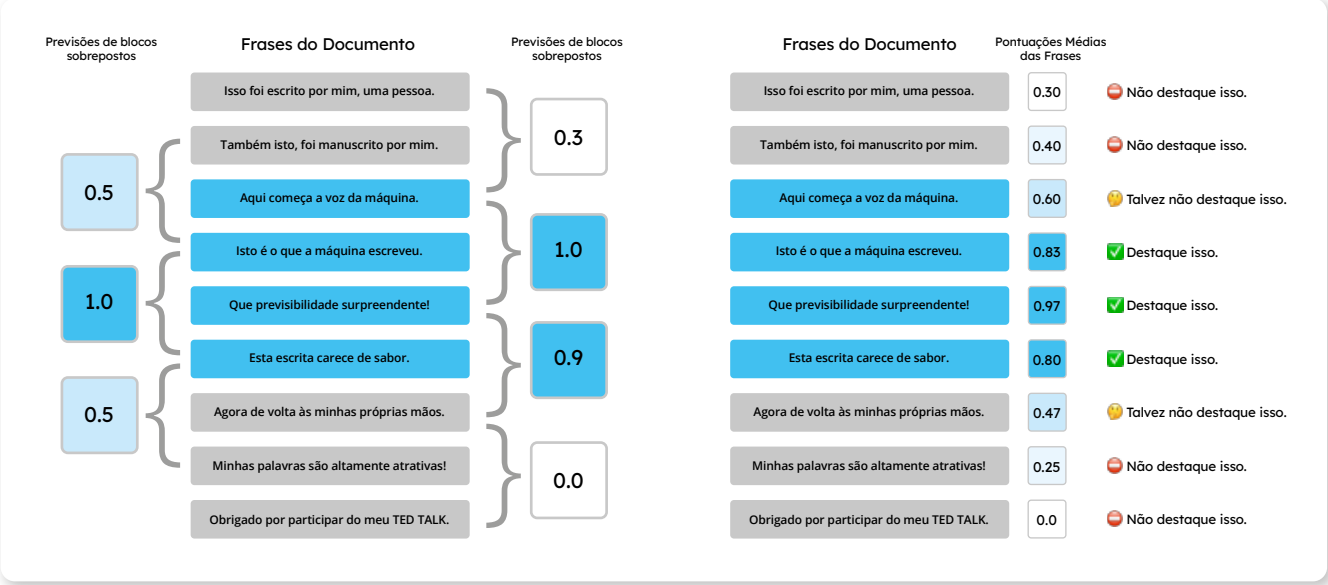
Sim. A Turnitin lançou seus recursos de detecção de escrita por IA para ajudar os professores a manterem a Integridade Acadêmica e, ao mesmo tempo, garantir que os alunos sejam tratados de forma justa.

Adicionamos um indicador de escrita por IA ao nosso Relatório de Similaridade, que mostra uma porcentagem geral do documento que as ferramentas de escrita por IA, como o ChatGPT, podem ter gerado. O indicador ainda está vinculado a um relatório que destaca os segmentos de texto que nosso modelo prevê serem possivelmente escritos por IA. Observe que apenas professores e administradores podem visualizar o indicador.

Embora a Turnitin tenha confiança em seu modelo, não fazemos uma determinação de má conduta, mas fornecemos dados para que os professores tomem uma decisão pautada e com base em suas políticas acadêmicas e institucionais. Portanto, devemos enfatizar que a porcentagem no indicador de escrita por IA não deve ser usada como base única para ação, ou medida definitiva de avaliação pelos professores.

2. Como funciona a solução?

Quando um artigo é submetido para a Turnitin, o envio é primeiramente dividido em segmentos de texto com aproximadamente algumas centenas de palavras (cerca de cinco a dez frases). Esses segmentos são então sobrepostos uns aos outros, para capturar cada frase no contexto.



Os segmentos são executados em nosso modelo de detecção de IA e damos a cada frase uma pontuação entre 0 e 1 para determinar se foi escrita por um humano, ou por IA. Se o nosso modelo determinar que uma sentença não foi gerada por IA, receberá uma pontuação de 0. Se determinar que toda a sentença foi gerada por IA, receberá uma pontuação de 1.

Utilizando as pontuações médias de todos os segmentos do documento, o modelo gera uma previsão geral de quanto do texto (com 98% de confiança com base nos dados coletados e verificados em nosso laboratório de inovação em IA) enviado acreditamos ter sido gerado por IA. Por exemplo, quando dizemos que 40% do texto geral foi gerado por IA, temos 98% de certeza de que é esse o caso.

Atualmente, o modelo de detecção de escrita por IA da Turnitin é treinado para identificar conteúdo dos modelos de linguagem GPT-3 e GPT-3.5, que inclui o ChatGPT. Estamos trabalhando ativamente na expansão de nosso modelo para nos permitir detectar melhor o conteúdo de outros modelos de linguagem de IA.

3. Quais parâmetros ou sinalizadores o modelo da Turnitin leva em consideração ao detectar a escrita por IA?

O GPT-3 e o ChatGPT são treinados com o texto de toda a Internet e, essencialmente, pegam essa grande quantidade de texto e geram sequências de palavras com base na escolha das próximas palavras altamente prováveis. Isso significa que o GPT-3 e o ChatGPT tendem a gerar o próximo verbete em uma sequência de palavras de maneira consistente e altamente provável. A escrita humana, por outro lado, tende a ser inconsistente e idiossincrática, resultando em uma baixa probabilidade de escolher a próxima palavra que o humano usará na sequência.

Nossos métodos de classificação são treinados para detectar essas diferenças na probabilidade de palavras e são adeptos das sequências de probabilidade de verbetes específicas de escritores humanos.

4. Como o modelo da Turnitin foi treinado?

Nosso modelo é treinado em uma amostra representativa de dados que inclui tanto a escrita acadêmica autêntica como aquela gerada por IA. Ao criar nosso conjunto de dados de amostra, levamos em consideração grupos estatisticamente sub-representados, como alunos de segunda língua, usuários de inglês de países que não tem o idioma como sua primeira língua, estudantes de faculdades e universidades de matrículas diversificadas e áreas menos comuns, como Antropologia, Geologia, Sociologia, entre outros.

5. Posso verificar envios de tarefas passadas para averiguar a escrita por IA?

Sim. Tarefas enviadas anteriormente podem ser verificadas em relação a detecção de escrita por AI se forem reenviadas para a Turnitin. Somente as tarefas enviadas após o lançamento de nosso recurso (4 de abril de 2023) são verificadas automaticamente para detecção de escrita por AI.

6. Quais idiomas são abordados?

Inglês. Para a primeira iteração dos recursos de detecção de escrita por IA da Turnitin, podemos detectar a escrita por IA para documentos enviados apenas em inglês, de formato longo.

7. O que acontecerá se um artigo em outro idioma for enviado?

Se for enviado um artigo que não esteja em inglês, o detector não processará o envio. O indicador mostrará um status de vazio/erro com orientação 'no aplicativo' que informará aos usuários que esse recurso funciona apenas para envios em inglês, no momento. Nenhum relatório será gerado se o conteúdo enviado não estiver em inglês.

8. Minha instituição pode obter acesso antecipado para poder testar esse novo recurso?

Não. Ao contrário dos lançamentos usuais de produtos da Turnitin, não podemos fornecer acesso antecipado para fins de teste a nenhum cliente para esta versão, pois estamos trazendo essa tecnologia para o mercado em uma velocidade acelerada, com base no feedback do cliente. No entanto, temos testado rigorosamente nossa tecnologia de detecção de IA em nossos laboratórios e estamos confiantes nos resultados.

Para ajudar os clientes a entenderem esse recurso, forneceremos documentação e orientação passo a passo sobre como usá-lo.

9. Eu ou meu administrador podemos suprimir o novo indicador e relatar se não quisermos vê-lo?

Não. Para este primeiro processo, não podemos suprimir o indicador de detecção de escrita por IA, ou o relatório, para qualquer cliente usando Turnitin Feedback Studio (TFS), TFS com Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check e Originality Check+ .

Recebemos uma quantidade significativa de feedbacks positivos dos clientes para avançar rapidamente com nossa tecnologia de detecção de escrita por IA, a fim de criar visibilidade quando se trata de alunos que usam este tipo de recurso de escrita e fornecer insights aos professores o mais rápido possível. Habilitar a supressão aumentaria nosso tempo de lançamento no mercado.

Para ajudar as instituições a aproveitarem ao máximo nossos recursos de detecção de IA, o recurso terá orientação no produto, bem como um link para uma página de perguntas frequentes explicando o significado dos vários resultados e como o recurso funciona. Fale com seu gerente de conta para obter acesso ao conjunto de slides explicando o funcionamento da ferramenta.

Como sempre, temos vários **recursos pedagógicos** disponíveis para os professores para ajudá-los a entender como manter a Integridade Acadêmica na era da IA.

10. A adição da funcionalidade de detecção de IA da Turnitin ao relatório de Similaridade mudará meu fluxo de trabalho, ou a maneira como uso este recurso?

Não. Essa funcionalidade adicional não altera a maneira como você usa o Relatório de Similaridade, ou seus fluxos de trabalho já existentes. Nossos recursos de detecção de IA foram adicionados ao Relatório de Similaridade para fornecer uma experiência ainda mais completa para nossos clientes.

11. Os recursos de detecção de IA estarão disponíveis via LMS (Sistema de Gestão de Aprendizagem), como Moodle, Blackboard, Canvas, etc? E o Microsoft Teams?

Sim, os usuários poderão ver o indicador e o relatório por meio do LMS que estiverem usando. Disponibilizamos a detecção de escrita por IA por meio do Relatório de Similaridade. Não há indicador de escrita por IA ou pontuação incorporada diretamente na interface do usuário do LMS, e os usuários precisarão acessar o relatório para visualizar a pontuação de IA.

Observe, no entanto, que a detecção de escrita por IA não estará disponível por meio da integração do Microsoft Teams. Isso ocorre porque a integração atual do MS Teams usa apenas o visualizador do aluno. Não há visualizador somente de professor separadamente. E, como nosso recurso de detecção de IA estará disponível apenas para professores, não podemos fornecê-lo por meio do MS Teams.

12. Como a detecção de autoria no Originality é diferente da detecção de escrita por IA?

A tecnologia de detecção de escrita por IA da Turnitin é diferente da tecnologia usada para criação (Originality). Nosso modelo de detecção de escrita por IA calcula a porcentagem geral de texto no documento enviado que provavelmente foi gerado por uma ferramenta de IA. A autoria, por outro lado, usa metadados, bem como análise de linguagem forense para detectar se a tarefa enviada foi escrita por alguém que não seja o aluno. Não poderá indicar se foi escrito por IA; só que o conteúdo não é obra do aluno.

Resultados de detecção de IA e interpretação

1. O que significa a porcentagem no indicador de detecção de escrita por IA?

A porcentagem indica a quantidade de texto qualificado no envio em que o modelo de detecção de escrita por IA da Turnitin identifica ser gerado por IA (com 98% de confiança com base em dados que foram cuidadosamente coletados e verificados em um ambiente de laboratório controlado). Este texto de qualificação inclui apenas sentenças em prosa, o que significa que analisamos apenas blocos de texto que são escritos em sentenças gramaticais padrão, e não incluem outros tipos de escrita, como listas, marcadores ou outras estruturas que não sejam sentenças.

Essa porcentagem não é necessariamente referente a todo o conteúdo enviado. Se o texto da submissão não for considerado texto em prosa de formato longo, ele não será incluído.

2. A porcentagem mostrada às vezes não corresponde à quantidade de texto destacado. Por quê?

Ao contrário do nosso Relatório de Similaridade, a porcentagem de escrita por IA não se correlaciona necessariamente com a quantidade de texto, no envio. O modelo de detecção de escrita por IA da Turnitin procura apenas frases em prosa contidas na escrita longa. O texto em prosa contido na escrita de formato longo significa frases individuais que compõem um trabalho escrito mais longo, como um ensaio, uma dissertação, um artigo, etc. O modelo não detecta texto gerado por IA, como poesia, scripts, ou código. Também não detecta escrita curta/não convencional, como marcadores, tabelas ou respostas curtas de exames.

3. Qual é a precisão do indicador de escrita por IA da Turnitin?

Nós somente sinalizamos algo como escrito por IA quando temos 98% de certeza de que foi gerado por IA. Isso porque queremos ter a segurança de não sinalizar falsamente algo que não foi gerado por IA. Significa, no entanto, que provavelmente perderemos até 15% do texto escrito por IA, com uma taxa de falsos positivos inferior a 1% (identificando incorretamente texto totalmente escrito por humanos como gerado por IA).

Por exemplo, se identificarmos que 50% de um documento é escrito por IA, temos 98% de certeza de que pelo menos 50% foi escrito por IA, com menos de 1% de taxa de falsos positivos, mas o texto pode conter até 65% de escrita gerada por IA.

As taxas acima foram determinadas por nosso modelo usando dados coletados e verificados em nosso AI Innovation Lab, mas sabemos que o uso no mundo real será diferente dos testes de laboratório. Para levar isso em consideração, ajustamos nosso detector de IA para minimizar falsos positivos em textos autênticos, mesmo que isso signifique que podemos perder algumas instâncias de escrita por IA.

4. O que posso fazer se achar que o indicador de IA está incorreto? Como o indicador da Turnitin lida com falsos positivos?

Se você encontrar documentos escritos por IA que não identificamos, ou perceber trabalhos autênticos de alunos que previmos como gerados por IA, informe-nos! Seu feedback é fundamental para nos ajudar a melhorar ainda mais nossa tecnologia. Você pode compartilhar seus comentários por meio do botão 'feedback' encontrado no relatório de escrita por IA.

Às vezes, falsos positivos (sinalizar incorretamente texto escrito por humanos como gerado por IA) podem incluir listas sem muita variação estrutural, texto que literalmente se repete, ou texto que foi parafraseado sem desenvolver novas ideias. Se nosso indicador mostrar uma quantidade maior de escrita por IA em tal texto, aconselhamos que você leve isso em consideração ao observar a porcentagem indicada.

Em um documento mais longo com uma mistura de escrita autêntica e texto gerado por IA, pode ser difícil determinar exatamente onde começa a escrita por IA e onde termina a escrita original, mas nosso modelo deve fornecer um guia confiável para iniciar conversas com o aluno responsável pelo envio.

Em documentos mais curtos, onde há apenas algumas centenas de palavras, a previsão será principalmente "tudo ou nada", porque estamos prevendo em um único segmento sem a oportunidade de sobreposição. Isso significa que algum texto que é uma mistura de conteúdo original e gerado por IA pode ser sinalizado como totalmente gerado por IA.

Por favor, considere estes pontos ao revisar os dados e abordar os alunos, ou outras pessoas.

5. Os alunos poderão ver os resultados?

O indicador e o relatório de detecção de escrita por IA não ficam visíveis para os alunos.

6. Qual é a diferença entre a pontuação de similaridade e a porcentagem de detecção de escrita por IA? Os dois são completamente separados ou eles sofrem influências um do outro?

A pontuação de similaridade e a porcentagem de detecção de escrita por IA são completamente independentes e não se influenciam. A **pontuação de similaridade** indica a porcentagem de correspondência de texto encontrada no documento enviado, quando comparada à coleção abrangente de conteúdo da Turnitin para verificação de similaridade.

A porcentagem de detecção de escrita por IA, por outro lado, mostra a porcentagem geral de texto em um envio que o modelo de detecção de escrita por IA da Turnitin prevê que foi gerado por ferramentas de escrita por IA.

7. O modelo da Turnitin leva em consideração que a tecnologia de detecção de escrita por IA pode ser tendenciosa contra áreas específicas, ou escritores de uma segunda língua?

Sim. Um dos princípios orientadores de nossa empresa e de nossa equipe de IA tem sido minimizar o risco de danos aos alunos, especialmente aqueles desfavorecidos ou privados de direitos pela história e estrutura de nossa sociedade. Portanto, ao criar nosso conjunto de dados de amostra, levamos em consideração grupos estatisticamente sub-representados, como alunos que possuem inglês como seu segundo idioma, usuários de inglês de países não majoritariamente nativos em inglês, alunos de faculdades e universidades historicamente desfavorecidas e áreas do conhecimento menos comuns, como Antropologia, Geologia, Sociologia e outras.

8. Como posso usar a porcentagem do indicador de IA em sala de aula com os alunos?

O indicador de detecção de IA da Turnitin mostra a porcentagem de texto que provavelmente foi gerada por uma ferramenta de escrita por IA, enquanto o relatório destaca os segmentos exatos que parecem ter sido escritos por IA. A decisão final sobre a ocorrência de qualquer má conduta cabe ao revisor/professor. A Turnitin não faz uma determinação de má conduta, mas fornece dados para que os professores tomem uma decisão pautada com base em suas políticas acadêmicas e institucionais.

Escopo de detecção

1. Quais modelos de escrita por IA a tecnologia da Turnitin pode detectar?

A primeira iteração dos recursos de detecção de escrita por IA da Turnitin foi treinada para detectar modelos incluindo GPT-3, GPT-3.5 e variantes. Nossa tecnologia também pode detectar outras ferramentas de escrita por IA baseadas nesses modelos, como o ChatGPT. Planejamos expandir nossos recursos de detecção para outros modelos no futuro.

2. O modelo atual é capaz de detectar texto gerado por GPT-4?

Estamos trabalhando constantemente para melhorar e expandir nossos recursos de detecção de escrita por IA. Atualmente, nossa equipe de IA está realizando testes com a tecnologia GPT-4 em nosso detector existente para comparar seu desempenho e entender as diferenças entre o GPT-3.5 (no qual nosso modelo é treinado) e o GPT-4. Os resultados preliminares são promissores, pois estamos detectando com precisão textos gerados por IA. Nossa análise está em andamento e, assim que estabelecermos métricas de eficácia confiáveis, atualizaremos nossos modelos para incluir o GPT-4. É importante observar, no entanto, que a versão gratuita do ChatGPT ainda está operando no GPT-3.5.

3. Como a Turnitin ficará imune ao futuro para versões avançadas do GPT e outros grandes modelos de linguagem que ainda surgirão?

Reconhecemos que os *Large Language Models* (LLM's - Grandes Modelos de Linguagem) estão se expandindo e evoluindo rapidamente, e já estamos trabalhando arduamente na criação de sistemas de detecção para LLMs adicionais. Nosso foco inicial é construir e lançar um detector de escrita por IA eficaz e confiável para GPT-3 e GPT-3.5, e outras ferramentas de escrita baseadas nesses modelos, como o ChatGPT.

4. A Turnitin pode detectar se o texto gerado por uma ferramenta de escrita por IA (ChatGPT, etc.) for posteriormente parafraseado usando uma ferramenta de paráfrase? Será sinalizado o conteúdo como gerado por IA mesmo neste caso?

Nossos detectores são treinados nas saídas de GPT-3, GPT-3.5 e ChatGPT, e modificar o texto gerado por esses sistemas terá um impacto na capacidade de nossos detectores de identificar texto escrito por IA. Em nosso AI Innovation Lab, realizamos testes usando ferramentas de paráfrase de código aberto (incluindo diferentes LLMs) e, na maioria dos casos, nosso detector manteve sua eficácia e é capaz de identificar o texto como gerado por IA, mesmo quando uma ferramenta de paráfrase foi usada para alterar a produção da IA.

5. A Turnitin tem planos de criar uma solução para detectar quando os alunos parafraseiam o conteúdo por conta própria ou por meio de ferramentas como o Quillbot etc.?

A Turnitin vem trabalhando na construção de recursos de detecção de paráfrase – capacidade de detectar quando os alunos parafraseiam o conteúdo com a ajuda de ferramentas de paráfrase ou o reescrevem – há algum tempo, e a tecnologia já está produzindo os resultados desejados em nosso *AI Innovation Lab*. No caso em que o aluno está usando um spinner de palavras ou uma ferramenta de paráfrase on-line, ele está apenas executando o conteúdo por meio de um recurso que usa IA para subverter intencionalmente a detecção de similaridade, não usando ferramentas generativas de IA, como o ChatGPT, para criar conteúdo.

Temos planos para uma versão beta em 2023 e disponibilizaremos a detecção de paráfrase para professores em instituições que usam o TFS com Originality e Originality por um custo adicional. A funcionalidade será lançada primeiro em nosso produto TFS com Originality.

Acesso e licenciamento

1. Quem terá acesso a esta solução? Teremos que pagar mais por essa funcionalidade?

A primeira iteração de nosso indicador e relatório de detecção de escrita por IA está disponível para nossos clientes de integridade de escrita acadêmica como parte de suas licenças já existentes, para que possam testar a solução e verificar como ela funciona. Isso inclui clientes com uma licença para Turnitin Feedback Studio (TFS), TFS com Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check e Originality Check+. O recurso está disponível para clientes que usam essas plataformas por meio de uma integração com um LMS ou com a API principal da Turnitin. Observe que apenas professores e administradores poderão ver o indicador e o relatório.

A partir de 1º de janeiro de 2024, apenas os clientes que licenciam o Originality ou o TFS com Originality terão acesso à experiência completa de detecção de escrita por IA.

2. Quando os clientes podem ter acesso a esta solução?

Os recursos de detecção de escrita por IA da Turnitin já estão disponíveis desde 04 de abril de 2023, e foram adicionados ao Relatório de Similaridade. Os clientes que licenciam qualquer um dos produtos Turnitin acima podem visualizar o indicador e acessar o relatório de IA.

3. A detecção de escrita por IA da Turnitin é uma solução independente ou faz parte de outro produto?

O primeiro contato com os recursos de detecção de escrita por IA da Turnitin é um recurso separado do Relatório de Similaridade e está disponível nos seguintes produtos: Turnitin Feedback Studio (TFS), TFS com Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check e Originality Check+. O indicador é vinculado a um relatório que mostra os segmentos exatos previstos como escritos por IA no conteúdo enviado.

4. Por que a detecção de IA não está sendo adicionada a outros produtos Turnitin, como Gradescope e iThenticate?

Concentramos nossos recursos no que visualizamos como o maior e mais intenso problema, que é direcionado ao Ensino Superior e à escrita longa no Ensino Fundamental e Médio. No momento, estamos investigando como podemos trazer a detecção de escrita por IA para os clientes do iThenticate. Atualmente, não temos planos de adicionar esses recursos ao Gradescope, pois o principal item de uso do Gradescope é o texto manuscrito, enquanto para a detecção de IA estamos nos concentrando no texto digitado. No entanto, estamos felizes em saber mais sobre as necessidades dos clientes para detecção de escrita por IA neste produto. Além disso, não estamos buscando detecção de código ChatGPT no momento.

5. Onde posso encontrar mais informações sobre esta nova solução?

Você pode encontrar informações sobre os recursos de detecção de escrita por IA da Turnitin **nesta página**.

101385 TII_GL_AI-GPT-4_BR_PT_0323