

OpenAI OpCo, LLC
Estatuto Digital da Criança e do Adolescente
(ECA Digital)
Relatório de Transparência

Lei nº 15.211/2025 (ECA Digital)

Período do relatório: jan. 1, 2026 a jun. 30, 2026

Índice

Índice.....	2
Sumário Executivo.....	4
Objetivo deste Relatório.....	4
Período e Status do Relatório.....	4
Escopo do Relatório.....	4
Compromisso com a Segurança, Privacidade e Transparência das Crianças e Adolescentes.....	4
1. Contexto Legal e Regulatório.....	4
2. Serviços Abrangidos.....	4
Escopo Territorial.....	5
3. Como Ler Este Relatório de Transparência.....	5
Definições Principais.....	5
Notas Metodológicas.....	6
Comparabilidade e Limitações.....	6
4. Canais de Denúncia e Sistemas e Processos de Apuração.....	6
Canais de Denúncia Disponíveis.....	6
Admissão, Triagem e Investigação.....	7
Ações e Escalonamento.....	7
5. Denúncias Recebidas, Prosseguimento, Moderação de Conteúdo e de Contas.....	7
Status e Ressalvas sobre os Dados.....	8
Informações sobre os Dados.....	8
ChatGPT.....	9
Denúncias recebidas e informações sobre o prosseguimento.....	9
Moderação de conteúdo e de contas por tipo.....	9
ChatGPT Search.....	10
Denúncias recebidas e informações sobre o prosseguimento.....	10
Moderação de conteúdo e de contas por tipo.....	10
GPTs.....	11
Denúncias recebidas e informações sobre o prosseguimento.....	11
Moderação de conteúdo e de contas por tipo.....	12
ImageGen.....	12
Denúncias recebidas e informações sobre o prosseguimento.....	12
Moderação de conteúdo e de contas por tipo.....	13
ChatGPT Sites.....	13
Denúncias recebidas e informações sobre o prosseguimento.....	13
Moderação de conteúdo e de contas por tipo.....	14
6. Medidas para Identificar Atos Ilegais.....	14
7. Melhorias Técnicas para Proteção e Privacidade de Dados.....	15
8. Avaliação de Impacto e Gerenciamento de Risco.....	16
Métodos Utilizados na Avaliação.....	16
Resultados, Medidas de Mitigação e Monitoramento.....	16

Sumário Executivo

Objetivo deste Relatório

Este Relatório de Transparência foi elaborado pela OpenAI OpCo, LLC (“**OpenAI**”) em conformidade com o art. 31 da Lei nº 15.211/2025 (Estatuto Digital da Criança e do Adolescente - “**ECA Digital**”)¹²³, o art. 45 do Decreto nº 12.880/2026, que regulamenta o ECA Digital,⁴ e o Despacho Decisório CD/ANPD nº 122/2026, expedido pela Agência Nacional de Proteção de Dados (“**ANPD**”).⁵ Seu objetivo é fornecer, em um único documento, as informações quantitativas e qualitativas exigidas dos provedores abrangidos por esta legislação.

Período e Status do Relatório

Este Relatório de Transparência abrange informações e métricas referentes ao período de **jan. 1, 2026** a **jun. 30, 2026** (o “**Período do Relatório**” ou apenas “**Período**”). Este Período coincide com o primeiro semestre civil especificado pela ANPD. Embora a ECA Digital tenha entrado em vigor em **mar. 17, 2026**, este Relatório abrange todo o semestre, a fim de fornecer informações completas e consistentes para o Período.

Escopo do Relatório

O Relatório abrange os serviços identificados na Seção 3 e se aplica às atividades realizadas durante o Período do Relatório no Brasil.

Compromisso com a Segurança, Privacidade e Transparência das Crianças e Adolescentes

A OpenAI está comprometida com a transparência, a responsabilidade e a proteção dos direitos fundamentais, incluindo a proteção de crianças e adolescentes, a liberdade de expressão e o acesso à informação. A OpenAI pretende continuar refinando e aprimorando suas divulgações ao longo do tempo.

1. Contexto Legal e Regulatório

O art. 31 da ECA Digital exige que os provedores de aplicações de internet direcionados a crianças e adolescentes, ou de acesso provável por crianças e adolescentes, publiquem relatórios semestrais de transparência em português em seus sites quando este tiver mais de um milhão de usuários cadastrados nessa faixa etária com conexão à internet no Brasil.

2. Serviços Abrangidos

Este Relatório abrange o seguinte:

¹ Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm.

² Disponível em: <https://www.gov.br/anpd/pt-br/assuntos/eca-digital>.

³ Disponível em: <https://infograficos.camara.leg.br/eca-digital/>.

⁴ Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12880.htm.

⁵ Disponível em: <https://www.in.gov.br/web/dou/-/despacho-decisorio-cd/anpd-n-122/2026-724474113>.

- **ChatGPT** - Funcionalidade de *chatbot* que permite aos usuários interagir de forma conversacional com os modelos da OpenAI por meio de linguagem natural e comandos multimodais, recebendo respostas geradas pelo sistema. Os usuários podem usar o ChatGPT para uma ampla variedade de tarefas, como organizar e resumir informações, auxiliar em traduções, realizar conversas por voz, analisar ou gerar imagens, apoiar a criatividade e a geração de ideias, escrever código e realizar outras tarefas do dia a dia.
- **Geração de imagens no ChatGPT (ImageGen)** - Permite que os usuários solicitem ao modelo que gere imagens.
- **ChatGPT search** - Funcionalidade do ChatGPT que permite ao serviço pesquisar na *web*, fornecendo respostas que podem incluir links para fontes relevantes, para que os usuários possam consultar as informações subjacentes.
- **GPTs** - Versões do ChatGPT configuradas sem necessidade de programação (*no-code*), para finalidades específicas dentro do ChatGPT. Os criadores podem personalizá-las por meio de instruções, arquivos de conhecimento e capacidades selecionadas.
- **ChatGPT Sites** - Um recurso que permite aos usuários elegíveis criar, publicar, hospedar e manter sites ou aplicativos web leves a partir do ChatGPT/Codex.

Escopo Territorial

Os dados apresentados abrangem às atividades relacionadas aos usuários dos serviços localizados no Brasil.

3. Como Ler Este Relatório de Transparência

Definições Principais

- **Denúncia** - Uma denúncia, incluindo reclamações, alegando que um conteúdo, conduta ou uma conta pode ter violado ou estar violando a legislação aplicável ou as políticas da OpenAI, recebida por meio de um mecanismo integrado ao produto ou um canal de denúncia externo. Para todos os fins, essa definição contempla o termo “denúncia”, previsto no art. 31, inciso II, do ECA Digital, e o termo “notificações”, previsto no art. 45, inciso I, do Decreto nº 12.880/2026.
- **Categoria de notificação** - A categoria selecionada pelo denunciante ou atribuída durante o recebimento da denúncia. Uma revisão final poderá identificar uma categoria diferente de violação.
- **Dados proporcionais ou prosseguimento** - Uma denúncia é considerada como tendo recebido um “prosseguimento” quando resulta em contato posterior com o denunciante, em uma investigação e/ou em uma ação de moderação de conteúdo ou de conta. A simples revisão não é contabilizada para esses fins, a menos que haja algum prosseguimento com o denunciante. Os resultados são apresentados em quantidade dentro da categoria relevante da denúncia.
- **Moderação de conteúdo** - Uma ação que afeta a disponibilidade, a visibilidade ou a acessibilidade da informação.

- **Moderação de conta** - Uma ação que afeta uma conta ou sua capacidade de acessar ou usar um serviço.
- **Período do relatório** - de 1º de janeiro de 2026 a 30 de junho de 2026.

Notas Metodológicas

Os dados apresentados refletem a atividade durante o Período do Relatório e são compilados de acordo com as metodologias descritas no Relatório. Quando um campo do Relatório não for aplicável, ele será deixado em branco. Quando um campo é aplicável, mas nenhuma atividade ocorreu, o valor reportado é zero.

Comparabilidade e Limitações

Diferenças no *design* do serviço, comportamento do usuário, obrigações legais, canais de denúncia e escopo da denúncia significam que os dados podem não ser diretamente comparáveis entre serviços ou períodos de reporte. As denúncias provenientes de formulários *web* externos e as denúncias provenientes dos mecanismos integrados ao produto são medidos separadamente e combinados apenas pela soma dos totais de cada canal; portanto, uma questão que aparece em mais de um canal pode ser contabilizada em cada um deles. Alterações em produtos, políticas ou metodologias de medição podem resultar em variações ao longo do tempo. Os leitores devem interpretar os dados à luz desses fatores contextuais. A OpenAI poderá publicar versões atualizadas deste Relatório para corrigir erros ou esclarecer metodologias.

4. Canais de Denúncia e Sistemas e Processos de Apuração

Canais de Denúncia Disponíveis

A OpenAI oferece funcionalidades de denúncia integradas ao produto, quando disponíveis, e um formulário online para denúncia de conteúdo⁶, por meio do qual usuários e membros do público podem denunciar conteúdos ou atividades que possam violar a legislação aplicável ou as políticas da OpenAI. Em consonância com o art. 28 do ECA Digital, reforçado pelo art. 41 do Decreto nº 12.880/2026, esses canais estão preparados para receber denúncias relativas a violações aos direitos de crianças e adolescentes.

Esses canais permitem o envio de denúncias sobre conversas e respostas do ChatGPT, geração de imagens, resultados de pesquisa ou outros conteúdos e atividades dentro de GPTs e ChatGPT Sites. As denúncias devem incluir informações suficientes para identificar o conteúdo ou a atividade em questão e permitir sua análise. O Centro de Ajuda da OpenAI oferece orientações sobre os canais de denúncia disponíveis e como enviar uma, sendo facilmente acessível aos usuários e disponibilizado gratuitamente.⁷

⁶ Disponível em: <https://openai.com/form/report-content/>.

⁷ Disponível em: <https://help.openai.com/en/articles/10245791-reporting-content-in-chatgpt-and-openai-platforms>.

Admissão, Triagem e Investigação

As Políticas de Uso da OpenAI exigem que os usuários cumpram as leis aplicáveis e proíbem condutas que explorem, coloquem em risco ou prejudiquem crianças e adolescentes. As denúncias podem identificar possíveis violações da legislação brasileira, bem como das políticas da OpenAI.

A OpenAI utiliza sistemas automatizados e pessoal treinado para auxiliar na avaliação do conteúdo denunciado. As denúncias são priorizadas de acordo com a gravidade do dano relatado, incluindo maior prioridade para denúncias sobre material de abuso sexual infantil.⁸ A análise automatizada pode ajudar a categorizar tais denúncias e identificar possíveis violações de políticas⁹. Se a revisão automatizada sinalizar que a denúncia pode violar as políticas, essa será encaminhada para revisão humana. Revisores humanos avaliam o conteúdo relatado em relação às políticas aplicáveis e escalam, quando necessário, questões que exigem avaliação adicional, como no caso de discussões sobre a aplicabilidade da legislação local, para equipes especializadas. Caso o conteúdo denunciado seja considerado uma violação das políticas da OpenAI ou da legislação aplicável, os revisores humanos determinarão as medidas cabíveis. Os processos de revisão estão sujeitos a procedimentos documentados, padrões de implementação de políticas, treinamento de pessoal e verificações de garantia de qualidade.

Ações e Escalonamento

Após análise, a OpenAI poderá tomar as medidas cabíveis de acordo com a violação identificada e a funcionalidade do serviço em questão. Dependendo das circunstâncias, as medidas¹⁰ podem incluir advertências, restrições ao compartilhamento de conteúdo ou à distribuição GPT, remoção de conteúdo de interfaces compartilhadas ou públicas, restrições às funcionalidades ou ao acesso ao produto e suspensão ou desativação da conta. Questões que exigem avaliação especializada podem ser escaladas para investigação adicional. Os usuários podem recorrer das medidas por meio dos processos internos, quando disponíveis, o que pode ser feito através do nosso formulário de recurso.¹¹

5. Denúncias Recebidas, Prosseguimento, Moderação de Conteúdo e de Contas

Esta Seção fornece uma visão geral das métricas, produto por produto, durante o Período do Relatório, para os usuários dos serviços localizados no Brasil.

A **primeira tabela** apresenta o número de denúncias recebidas, o número de notificações por categoria, o prosseguimento dessas denúncias e a respectiva proporção representada, em conformidade com o art. 31, II do ECA Digital e o art. 45 do Decreto nº 12.880/2026.

⁸ Disponível em: <https://openai.com/index/combating-online-child-sexual-exploitation-abuse>.

⁹ Disponível em: <https://openai.com/policies/usage-policies/>; <https://openai.com/policies/row-terms-of-use/>; <https://openai.com/policies/service-terms/>

¹⁰ Disponível em: <https://openai.com/transparency-and-content-moderation/>

¹¹ Disponível em: <https://openai.com/pt-BR/form/appeal/>

Conforme detalhado acima, a OpenAI realiza uma avaliação individualizada do conteúdo denunciado por meio de ferramentas automatizadas e revisão humana antes de tomar qualquer medida punitiva. Nem todos os relatórios resultam em uma constatação de não conformidade. Em muitos casos, o conteúdo denunciado pelos usuários é considerado, após análise, consistente com as políticas da OpenAI e os requisitos legais aplicáveis, e nenhuma medida é tomada. Em outros casos, a denúncia coincide com um comportamento inadequado do usuário que desencadeia medidas sancionatórias ou administrativas, que podem incluir: **(1)** remoção ou restrição de conteúdo; **(2)** advertência; **(3)** restrição ou suspensão de conta; ou **(4)** encerramento de conta.

A **segunda tabela** abaixo apresenta os números relativos às ações de moderação por tipo, em conformidade com o art. 31, III do ECA Digital, utilizando uma estrutura baseada em produto por produto.

Status e Ressalvas sobre os Dados

As denúncias são contabilizadas uma única vez em cada canal; cada ação de moderação é contabilizada separadamente; e a aparente sobreposição entre produtos ou entre o formulário *web* de denúncia e o mecanismo integrado ao produto é sinalizada em vez de ser eliminada sem indicação expressa. Note que o formulário *web* começou a associar casos ao país de residência habitual do usuário apenas a partir da metade de março de 2026, portanto, seus números podem não representar a cobertura completa de todo o Período do Relatório.

Os valores relativos ao prosseguimento das denúncias nas colunas “Dados Proporcionais” abaixo correspondem à soma aritmética dos desfechos registrados nos mecanismos integrados ao produto e das ocorrências de envio de e-mails de acompanhamento referentes às denúncias pelo formulário *web*. Esses valores não constituem contagens realizadas de denúncias distintas que receberam prosseguimento. Não foram calculados percentuais..

Informações sobre os Dados

As seguintes informações devem ser levadas em consideração sobre as categorias:

- “Exploração e abuso sexual infantil” inclui denúncias de atividades de sexualização infantil, material de abuso sexual infantil (*child sexual abuse material* ou “**CSAM**”) e aliciamento.
- “Crimes contra a democracia e o Estado de Direito” inclui denúncias de interferência eleitoral e infrações legais.
- “Incentivo ou promoção à violência grave” inclui denúncias de incitação à violência, incluindo à violência sexual.
- “Indução ou assistência ao suicídio ou à automutilação” inclui denúncias relacionadas à promoção ou facilitação de transtornos alimentares e de autolesão não suicida.
- “Exploração sexual” inclui denúncias de facilitação de violência sexual, conteúdo íntimo não consensual e sexualização de pessoas reais.
- OpenAI pode adotar mais de uma ação em resposta a um único problema; consequentemente, o número de ações pode exceder o número de denúncias ou até mesmo de itens de conteúdo ou contas analisadas.

ChatGPT

Denúncias recebidas e informações sobre o prosseguimento

Categoria da denúncia	Denúncias recebidas	Dados proporcionais
Exploração e abuso sexual infantil	384	190
Conteúdo sobre <i>bullying</i>	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Assédio, ameaças e abuso	1.344	520
Substâncias e drogas proibidas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Uso de substâncias nocivas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo relacionado ao suicídio	1	0
Conteúdo pornográfico	122	28
Terrorismo	88	36
Indução ou assistência ao suicídio ou à automutilação	135	51
Discurso de ódio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Tráfico de pessoas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Crimes contra a democracia e o Estado de Direito	2.091	739
Incitação ao ódio e à discriminação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incentivo ou promoção de violência grave	635	258
Violência grave realista e lesões gráficas	136	47
Exploração sexual	1.317	616

Moderação de conteúdo e de contas por tipo

Tipo de moderação	Contagem
Remoção ou restrição de conteúdo	2.436
Advertência	9
Restrição ou suspensão da conta	0
Encerramento da conta	2
Outros	0
Número total de ações	2.447

ChatGPT Search

Denúncias recebidas e informações sobre o prosseguimento

Categoria da denúncia	Denúncias recebidas	Dados proporcionais
Exploração e abuso sexual infantil	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre <i>bullying</i>	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Assédio, ameaças e abuso	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Substâncias e drogas proibidas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Uso de substâncias nocivas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo relacionado ao suicídio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo pornográfico	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Terrorismo	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Indução ou assistência ao suicídio ou à automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Discurso de ódio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Tráfico de pessoas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Crimes contra a democracia e o Estado de Direito	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incitação ao ódio e à discriminação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incentivo ou promoção à violência grave	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Violência grave realista e lesões gráficas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Exploração sexual	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	

Moderação de conteúdo e de contas por tipo

Tipo de moderação	Contagem
Remoção ou restrição de conteúdo	N/A
Advertência	N/A

Tipo de moderação	Contagem
Restrição ou suspensão da conta	N/A
Encerramento da conta	N/A
Outros	N/A
Número total de ações	N/A

GPTs

Denúncias recebidas e informações sobre o prosseguimento

Categoria da denúncia	Denúncias recebidas	Dados proporcionais
Exploração e abuso sexual infantil	2	1
Conteúdo sobre <i>bullying</i>	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Assédio, ameaças e abuso	9	9
Substâncias e drogas proibidas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Uso de substâncias nocivas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo relacionado ao suicídio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo pornográfico	3	3
Terrorismo	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Indução ou assistência ao suicídio ou à automutilação	1	1
Discurso de ódio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Tráfico de pessoas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Crimes contra a democracia e o Estado de Direito	8	7
Incitação ao ódio e à discriminação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incentivo ou promoção à violência grave	4	3
Violência grave realista e lesões gráficas	1	1
Exploração sexual	6	6

Moderação de conteúdo e de contas por tipo

Tipo de moderação	Contagem
Remoção ou restrição de conteúdo	30
Advertência	0
Restrição ou suspensão da conta	0
Encerramento da conta	0
Outros	0
Número total de ações	30

[ImageGen](#)

Denúncias recebidas e informações sobre o prosseguimento

Categoria de notificação	Denúncias recebidas	Dados proporcionais
Exploração e abuso sexual infantil	1	0
Conteúdo sobre <i>bullying</i>	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Assédio, ameaças e abuso	5	0
Substâncias e drogas proibidas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Uso de substâncias nocivas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo relacionado ao suicídio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo pornográfico	1	0
Terrorismo	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Indução ou assistência ao suicídio ou à automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Discurso de ódio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Tráfico de pessoas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Crimes contra a democracia e o Estado de Direito	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incitação ao ódio e à discriminação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incentivo ou promoção à violência grave	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	

Categoria de notificação	Denúncias recebidas	Dados proporcionais
Violência grave realista e lesões gráficas	1	0
Exploração sexual	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	

Moderação de conteúdo e de contas por tipo

Tipo de moderação	Contagem
Remoção ou restrição de conteúdo	N/A
Advertência	N/A
Restrição ou suspensão da conta	N/A
Encerramento da conta	N/A
Outros	N/A
Número total de ações	N/A

ChatGPT Sites

Denúncias recebidas e informações sobre o prosseguimento

Categoria da denúncia	Denúncias recebidas	Dados proporcionais
Exploração e abuso sexual infantil	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre <i>bullying</i>	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Assédio, ameaças e abuso	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Substâncias e drogas proibidas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Uso de substâncias nocivas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo sobre automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo relacionado ao suicídio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Conteúdo pornográfico	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Terrorismo	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Indução ou assistência ao suicídio ou a automutilação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Discurso de ódio	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Tráfico de pessoas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	

Categoria da denúncia	Denúncias recebidas	Dados proporcionais
	do Relatório.	
Crimes contra a democracia e o Estado de Direito	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incitação ao ódio e à discriminação	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Incentivo ou promoção à violência grave	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Violência grave realista e lesões gráficas	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	
Exploração sexual	N/A - Não foram identificadas denúncias para esta categoria durante o Período do Relatório.	

Moderação de conteúdo e de contas por tipo

Tipo de moderação	Contagem
Remoção ou restrição de conteúdo	N/A
Advertência	N/A
Restrição ou suspensão da conta	N/A
Encerramento da conta	N/A
Outros	N/A
Número total de ações	N/A

6. Medidas para Identificar Atos Ilegais

O art. 31, IV do ECA Digital estabelece que os provedores de aplicações de internet devem divulgar as medidas adotadas para identificar atos ilícitos. A OpenAI utiliza diversos procedimentos e ferramentas para proteger os seus usuários, incluindo os adolescentes, de conteúdo ilegal e nocivo. As Políticas de Uso definem claramente os tipos de conteúdo que não são permitidos nos serviços da OpenAI, inclui¹²ndo conteúdo ilegal ou nocivo a crianças e adolescentes. Uma série de medidas foi implementada para prevenir, detectar e responder a violações, incluindo ações de fiscalização contra conteúdo que viola as regras aplicáveis.¹³ Nesse sentido, a OpenAI apresenta abaixo uma visão geral das principais medidas.

Mecanismos de Aferição de Idade

A OpenAI utiliza um modelo de estimativa de idade¹⁴ que nos ajuda a reforçar a proteção de usuários menores de 18 anos e aplicar as medidas necessárias quando uma conta pode pertencer a um usuário menor de 18 anos. Para usuários de 13 a 17 anos, o ChatGPT aplica proteções

¹² Disponível em: <https://openai.com/pt-BR/policies/usage-policies/>.

¹³ Disponível em: <https://openai.com/pt-BR/policies/brazil-eca-digital/>.

¹⁴ Disponível em: <https://help.openai.com/en/articles/12652064-age-prediction-in-chatgpt>.

adicionais¹⁵ projetadas para bloquear conteúdo sensível ou potencialmente nocivo, incluindo automutilação, violência, transtornos alimentares, atividades perigosas e conteúdo sexual ou gráfico explícito. Quando nosso modelo não tem certeza da idade de um usuário, incluindo, por exemplo, quando um usuário é novo no serviço, o usuário será direcionado para uma experiência mais segura com medidas de proteção.

Medidas Proativas para Identificação de Conteúdo Ilegal

A OpenAI utiliza medidas técnicas e organizacionais em camadas para identificar e responder a conteúdos potencialmente ilegais ou que violem as políticas da OpenAI e que afetem crianças e adolescentes. Essas medidas incluem avaliação automatizada de *prompts* e respostas, verificações de segurança de imagem e revisão adicional assistida por modelo ou humana.¹⁶ As salvaguardas contra abuso sexual infantil incluem ferramentas para identificar material conhecido e potencialmente novo. Para o ChatGPT Search, as medidas relevantes também incluem classificadores de segurança (*web-safety classifiers*) e listas de bloqueio de URLs e domínios, projetadas para impedir que material proibido identificado apareça nos resultados da busca.

Dependendo do serviço e das circunstâncias relevantes, as violações identificadas podem resultar na recusa de uma solicitação, na retenção de uma resposta ou imagem, em restrições de compartilhamento ou acesso, ou na aplicação de medidas contra a conta.¹⁴ As denúncias e sinais de segurança podem ser encaminhados para avaliação especializada por meio dos processos descritos na Seção 5.

7. Melhorias Técnicas para Proteção e Privacidade de Dados

De acordo com o art. 31, V do ECA Digital, os provedores devem apresentar os aprimoramentos técnicos relativos à proteção e privacidade dos dados. Neste contexto, as medidas implementadas pela OpenAI incluem controles sobre usos selecionados de informações e medidas projetadas para reduzir divulgações nocivas por meio de seus serviços. Os controles de supervisão parental em contas de adolescentes vinculadas permitem que os pais ou responsáveis controlem a experiência dos adolescentes ao, por exemplo, desativar a funcionalidade de memória.¹⁷ A OpenAI também aplica políticas e classificadores de segurança projetados para restringir solicitações ou resultados que revelem dados pessoais altamente sensíveis ou que facilitem a localização, o rastreamento ou a intimidação de uma pessoa identificável. Dependendo da funcionalidade e das circunstâncias relevantes, o conteúdo potencialmente violador pode ser bloqueado ou encaminhado para análise posterior.

Durante o Período do Relatório, a OpenAI implementou a estimativa de idade no ChatGPT para aplicar salvaguardas adicionais a contas que provavelmente pertencem a usuários menores de 18

¹⁵ Disponível em: <https://openai.com/index/chatgpt-for-teens/>

¹⁶ Disponível em: <https://help.openai.com/en/articles/8940831-how-we-identify-problematic-content-on-our-services-for-individuals>

¹⁷ Disponível em: <https://help.openai.com/en/articles/12315553-parental-controls-faq>

anos — usuários que acreditam ter sido classificados incorretamente podem verificar sua idade por meio de nosso provedor terceirizado, a Persona. Implementamos lembretes "Faça uma pausa" baseados em tempo para usuários menores de 18 anos, que são acionados após 90 minutos de atividade em um período contínuo de 3 horas. A OpenAI também excluiu usuários adolescentes do acesso à publicidade e introduziu horários de silêncio para notificações, suprimindo-as para reduzir as interrupções noturnas.

Para promover um uso construtivo e adequado à idade, a OpenAI introduziu blocos de aprendizagem interativos e continuou a expandir sua abrangência. Essas medidas ajudam os usuários a explorar e compreender conceitos em disciplinas como matemática e ciências. No Brasil, a OpenAI também ampliou o acesso à educação em segurança familiar por meio de um guia de alfabetização em IA para famílias, adaptado ao contexto local.¹⁸ Um esforço de divulgação mais amplo inclui campanhas de conscientização explicando as medidas de segurança e os controles parentais disponíveis.

8. Avaliação de Impacto e Gerenciamento de Risco

Métodos Utilizados na Avaliação

A avaliação de risco da OpenAI no contexto do ECA Digital considera os riscos previsíveis para crianças e adolescentes decorrentes do *design* e operação normais do ChatGPT, incluindo o ChatGPT Search, ImageGen, os GPTs e o ChatGPT Sites. A avaliação examina como o conteúdo nocivo pode ser gerado, recuperado, encontrado ou disponibilizado por meio de cada funcionalidade, inclusive por meio de solicitações deliberadas ou exposição accidental. Essa avaliação considera os potenciais efeitos tanto em usuários quanto em não usuários, incluindo crianças ou adolescentes retratados ou visados pelo conteúdo, bem como as diferenças de vulnerabilidade associadas à idade, ao estágio de desenvolvimento e às circunstâncias individuais.

Para cada categoria de dano identificada, a OpenAI avalia qualitativamente a probabilidade de exposição, a natureza e a gravidade potencial do dano, as salvaguardas aplicáveis e a medida em que essas salvaguardas reduzem a probabilidade ou o impacto, à luz das evidências disponíveis. A avaliação atribui um nível de risco final de baixo, médio ou alto. A menor probabilidade de exposição se distingue da gravidade do dano caso ocorra um incidente.

A avaliação faz parte de um processo contínuo de segurança que incorpora avaliações pré-implantação, testes adversariais e *red teaming*, monitoramento pós-implantação, relatórios de usuários e revisão humana. Essas atividades contribuem para a avaliação e o aprimoramento do comportamento do modelo, das salvaguardas do produto e dos processos de fiscalização.

Resultados, Medidas de Mitigação e Monitoramento

A avaliação analisa as categorias relevantes de risco com base no marco legal brasileiro, incluindo aquelas relacionadas à exploração e abuso sexual infantil; *bullying*, assédio, ameaças e discriminação; drogas ilícitas e uso de substâncias nocivas; automutilação e suicídio; pornografia e exploração sexual; terrorismo e tráfico de pessoas; violência grave e lesões corporais graves; e

¹⁸ Disponível em: <https://openai.com/pt-BR/index/ai-literacy-resources-for-teens-and-parents/>.

condutas que afetam a democracia e o Estado de Direito e que podem causar danos à segurança da criança ou do adolescente.

Levando em consideração o *design* das funcionalidades avaliadas e as salvaguardas aplicáveis, a avaliação conclui que essas medidas reduzem materialmente a probabilidade de exposição ou facilitação dos riscos identificados. Apesar disso, ainda existem riscos residuais, cuja natureza e impacto potencial variam de acordo com a funcionalidade, o contexto de uso e as circunstâncias das crianças e adolescentes afetados. Uma probabilidade reduzida não diminui a gravidade potencial de um incidente individual. Esses riscos residuais continuam sujeitos à monitoramento, avaliação e aprimoramento contínuos das salvaguardas relevantes.

A OpenAI analisa sinais de segurança, denúncias de usuários, resultados de ações de *enforcement* e desempenho de controles para identificar riscos emergentes e orientar melhorias. As atividades relevantes incluem avaliações de segurança, testes adversariais, análise especializada de ameaças e revisão humana de conteúdo potencialmente violador. Os resultados podem orientar mudanças no treinamento de modelos, políticas, mecanismos de detecção e práticas de fiscalização.