Keeping Users Safe in the Age of AI

Our safety journey consists of four important steps – beginning with model pre-training (where the model learns patterns), then model post-training (where the model leverages learned patterns to provide helpful answers), next to pre-deployment evaluations (where the model goes through capability and safety evaluations as well as red teaming), and finally to post-deployment (where we monitor usage and take appropriate actions).

Safety at every step

We design and train our models to prioritize safety from the start

Responsible Model Training

We are careful what information we use to teach our AI. We avoid and filter child sexual abuse material (CSAM) and child sexual exploitation material (CSEM), and we report any attempts by users to upload it. We always stay vigilant and monitor for attempts to produce sexual content involving minors with our AI and block them.

Real-Time Detection

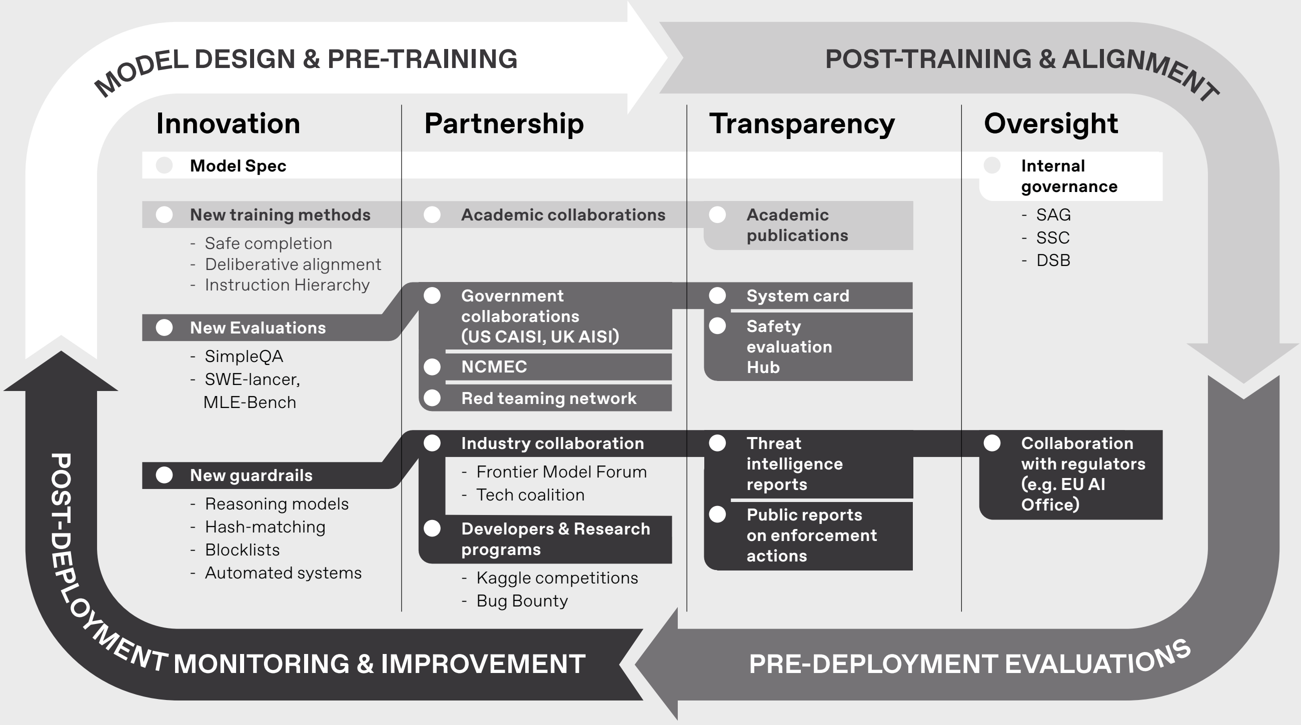
We combine automated systems and human review to identify and block harmful material. Our classifiers assess text, image, and audio content in real time. We use tools such as Thorn's Safer classifier and blocklists from partners like the Internet Watch Foundation to detect and report child sexual abuse material (CSAM).

Safeguarding Against Misuse and Harmful Content

OpenAI publishes product usage policies for ChatGPT and other services that prohibit harmful uses such as dangerous challenges for minors, promotion of suicide self harm or eating disorders, scams and more. These policies are enforced not just through terms and policies, but also through technical systems that help prevent and detect misuse.

Responding to Sensitive Conversations

We've been using our real-time router to direct sensitive parts of conversations — such as those showing signs of acute distress — to reasoning models. When GPT-5 Auto or a non-reasoning model is selected, we'll instead route these conversations to GPT-5 Instant to more quickly provide helpful and beneficial responses. These updates were guided by mental health experts and help ChatGPT de-escalate conversations and point people to real-world crisis resources when appropriate.



MODEL DESIGN & PRE-TRAINING

POST-TRAINING & ALIGNMENT

PRE-DEPLOYMENT EVALUATIONS

POST-DEPLOYMENT MONITORING & IMPROVEMENT

Innovation

- Model Spec
- New training methods
  - Safe completion
  - Deliberative alignment
  - Instruction Hierarchy
- New Evaluations
  - SimpleQA
  - SWE-lancer, MLE-Bench
- New guardrails
  - Reasoning models
  - Hash-matching
  - Blocklists
  - Automated systems

Partnership

- Academic collaborations
- Government collaborations (US CAISI, UK AISI)
- NCMEC
- Red teaming network
- Industry collaboration
  - Frontier Model Forum
  - Tech coalition
- Developers & Research programs
  - Kaggle competitions
  - Bug Bounty

Transparency

- Academic publications
- System card
- Safety evaluation Hub
- Threat intelligence reports
- Public reports on enforcement actions

Oversight

- Internal governance
  - SAG
  - SSC
  - DSB
- Collaboration with regulators (e.g. EU AI Office)

Safety & Responsibility at OpenAI

This reflects our views as of October 2025.
We will update this document as AI capabilities continue to advance.

Safety Evaluations Hub

Supporting young people and their families with safety systems, parental controls and learning environments

Applying age-appropriate experiences

Our goal is to create a safer experience for every user, by default

We're building a long-term system to predict whether a user is over or under 18, so we can **tailor their ChatGPT experience accordingly.**

When we identify a user as under 18, they are automatically directed to a more protected version of ChatGPT, which includes measures like reducing graphic content, viral challenges, sexual, romantic or violent roleplay, and extreme beauty ideals.

When we are not confident about an estimation of someone's age or have incomplete information, **we take the safer route** and default to the under-18 experience. Adults will also have secure ways to verify their age to access adult-specific features.

Set up parental controls

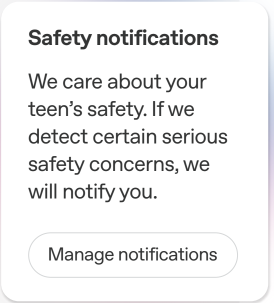
Parents can link their teen's ChatGPT account — with the teen's permission — to manage features and tailor how the system responds to their teen. Teens also have their own privacy rights that are respected and balanced within this setup

- **Disable features** like memory or chat history
- **Set "quiet hours"** when ChatGPT is unavailable
- **Receive alerts** if our systems detect signs of possible self harm

Safety notifications

By default, parents don't have access to their teen's conversation data.

In rare cases where our system and trained reviewers detect possible signs of serious self harm, parents may be notified. Even in these rare situations, we take teen privacy seriously and will only share the information needed for parents or emergency responders to protect a teen's safety.



Safety notifications

We care about your teen's safety. If we detect certain serious safety concerns, we will notify you.

Learning and creativity features

ChatGPT can support learning and exploration in safe, age-appropriate ways

- **Study Mode** helps students reason through problems step-by-step rather than just giving answers.
- **Teaching with AI** offers practical classroom guidance and prompt examples.
- **ChatGPT Foundations for Educators** developed with Common Sense Education to help teachers promote safe and creative use.

Parental Resource Hub

Partnering for Better Protection

Protecting young users online is a shared responsibility
We work across sectors to make AI development safer, more transparent, and more informed by evidence

Ground our approach in evidence
We collaborate with psychologists, educators, and civil-society organizations to translate research into real-world safety measures that guide product design and policy.

Collaborate with experts worldwide
Through our Well-Being Council and Global Physician Network—bringing together over 250 specialists from 60 countries—we draw on expertise in youth development, digital well-being, pediatrics, and mental health to shape safer AI systems.

Invest in open and transparent safety tools
As a founding member of ROOST, a nonprofit building accessible and transparent safety standards for technology platforms, we help advance collective solutions for online protection.

Bringing AI into Classrooms Safely
In Estonia and Greece ChatGPT is being rolled out in public schools as part of national programs to equip teachers with the skills to use AI safely and effectively in education.

This reflects our views as of October 2025. We will update this document as AI capabilities continue to advance.