Privacy Designed With Intelligence

AI Model

Pre-Training

Data Collection

We train our models primarily on publicly available data and data that we access through proprietary data partnerships, while taking steps to exclude sources that are access restricted, sites that primarily aggregate personal or sensitive information, or content that violates our policies

First Filtering

We exclude known sources like personal data aggregators and websites known for spam, illegal content, or harmful material to prevent the model from learning from such data

Deduplication

We take steps to identify and remove duplicate content—including personal data—to reduce the chance of the model learning information about private individuals

Tokenization (Unstructured Data)

Before training, we convert unstructured text into numerical tokens representing words or sub-word units, and do not index or link data to individual identities

Second Filtering

We apply a custom-built tool developed by our privacy engineering team to detect and mask a wide range of personal information that may be incidentally present in training data, reducing the likelihood the models learn from such information while preserving necessary context for utility

Synthetic Data

We increasingly use our own models to generate synthetic data for training future models, leveraging our existing models to continuously enhance privacy protections over time

At OpenAI we lead the way in designing and **embedding privacy-preserving** safeguards into **each step of the AI lifecycle** — from pre training to deployment

Post-Training

Reinforcement Learning From Human Feedback

We incorporate human feedback from AI trainers, red teamers, employees, and users whose data control settings allow model improvements

Red-Teaming

We collaborate with internal and external experts to rigorously test models for potential harms—including privacy risks—across diverse domains, feeding insights back into model improvements

ChatGPT

Deployment

User Control

We give users the ability to opt out of having their conversations used for training and allow them to request the removal of verifiable personal data from model outputs, reinforcing user agency and control

Notice & Transparency

We provide clear public explanations about how our models are trained and used, including our approach to privacy, and offer users access to policies and tools that explain and govern data use

Output Refusals

Our models are trained to decline requests for certain private or sensitive information, including data about individuals—even if publicly available

Usage Policies

We enforce strict usage policies and model behavior guidelines intended to prevent the generation of harmful, illegal, or privacy-violating outputs that violate our policies, and continually updates these safeguards as models evolve

Disconnecting From User Account

Before using ChatGPT conversations for model improvement, we disassociate them from user accounts by separating and filtering identifying information, so that training data is not linked to individuals

Safeguards to Minimize Reproducing Training Data

We post-train our models to mitigate models repeating training data

Refusals to Avoid Providing Private or Sensitive Info

We fine-tune our models to reject requests for private or sensitive personal information—even if that data exists online—to better protect individual privacy

This reflects our views as of August 2025.
We will update this document as AGI continues to evolve.

Learn More About Consumer Privacy at OpenAI

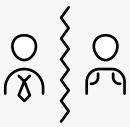
Advancing Privacy with AI

Protecting people's privacy **is core to building safe and trustworthy AI**. That's why OpenAI created **Privacy Filter** — a powerful privacy tool built using our most advanced language models to detect and mask personal information in training data.

What is Privacy Filter?

A privacy-first AI tool designed to:

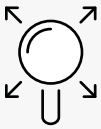
Detect a wide range of personal information like names, phone numbers, and emails—in large, messy datasets.



Distinguish between public and private context like separating an official business address from a private individual's contact details.

Mask personal information before training, helping to ensure that models don't learn from such information while preserving valuable context.

How is Privacy Filter different?



Broader detection
Goes far beyond standard tools available in the market to recognize a broader range of personal information.

Smarter Decisions
Understands the difference between a famous scientist's biography and a private individual's contact details.

Proven Results
Privacy Filter achieved state-of-the-art precision — outperforming industry leaders across benchmarking evaluations.

Built by Privacy, for Privacy

We trained Privacy Filter using our own models and a detailed privacy taxonomy. It adapts to new types of personal info—even those it's never seen before—while aligning closely with human judgment in real-world tests.

This reflects our views as of August 2025. We will update this document as AGI continues to evolve.