# Responsible Use of Anthropic's Models: Guidelines for Organizations Serving Minors

Updated over 2 weeks ago

At Anthropic, we recognize the unique vulnerabilities and needs of children in digital spaces. In order to create a safer digital environment and mitigate risks, organizations providing minors with the ability to directly interact with products that incorporate our API(s) should implement the following safeguards:

**1. Additional Technical Measures**

Organizations with products serving minors should implement additional safety features tailored to their unique use cases, as they are best situated to understand the specific ways their end users may interact with products that incorporate Anthropic's services. These safety measures may include, but are not limited to:

- Age verification systems to ensure only intended users can access the product
- Content moderation and filtering to block inappropriate or harmful content
- Monitoring and reporting mechanisms to identify and address potential issues
- Educational resources and guidance for minors on safe and responsible use of the product

In addition to these organization-specific measures, Anthropic may make available technical measures intended to tailor product experiences for certain end users, including minors. For example, we may provide a child-safety system prompt, which organizations serving minors should implement as part of a comprehensive suite of safety measures. It is important to note that, while helpful, these technical measures are not infallible and should be used in conjunction with the organization's own safety features to ensure a robust approach to child safety.

**2. Regulatory Compliance**

It is the responsibility of organizations to comply with all applicable child safety and data privacy regulations, such as the Children's Online Privacy Protection Act (COPPA) in the United States. Compliance with these regulations should be clearly stated on the organization’s website or similar public-facing documentation.

**3. Disclosure Requirements**

Organizations must disclose to their users that they are interacting with an AI system rather than a human.

Anthropic will periodically audit organizations for compliance with these safeguards. If your organization has a high violation rate and has not implemented these safety recommendations, we may ask you to implement them. Failure to implement these

recommendations when requested, or a continued high violation rate, may lead to the suspension or termination of your account.

Safeguards

Using Agents According to Our Usage Policy

# Using Agents According to Our Usage Policy

Updated over a week ago

All uses of agents and agentic features must continue to adhere to Anthropic's Usage Policy. The following are intended to be non-exhaustive illustrations of how our Usage Policy applies to certain agentic uses. As agentic capabilities evolve, we will update this list with additional examples to help users understand what our Usage Policy covers in practice in agentic environments.

## Do Not Use Agents for Surveillance or Unauthorized Data Collection

This includes using agents to:

- Monitor or track individuals' online activities, behaviors, or movements without notification or consent
- Collect, compile, or analyze personal information to create profiles based on individuals' protected attributes, sensitive characteristics, or personal circumstances
- Use facial recognition or biometric identification software or websites
- Conduct mass surveillance across multiple websites or platforms to send communications or engage in any form of targeted actions

## Do Not Use Agents to Generate or Distribute Harmful Content

This includes using agents to:

- Create websites or domains that mimic legitimate webpages
- Generate content that leads to phishing, social engineering, or fraud
- Impersonate individuals (private or public) without their consent

## Do Not Use Agents to Engage in Scaled Abuse

This includes using agents to:

- Spam government services, emergency systems, or crisis helplines
- Overwhelm servers with traffic to disrupt services (e.g., DDoS attacks)
- Coordinate harassment campaigns across multiple platforms or accounts
- Manipulate online polls, voting systems, or traffic metrics

- Create or manage multiple accounts to evade detection or circumvent platform safeguards
- Engage in click farming or artificial engagement (e.g., through likes or comments) on social media
- Automate influence operations or coordinated inauthentic behavior
- Bulk report people, users, or content through abuse reporting systems

### Do Not Use Agents for Unauthorized System Access or Manipulation

This includes using agents to:

- Install malware, backdoors, or monitoring software without authorization
- Execute commands that attempt privilege escalation or system exploitation
- Perform actions that could compromise critical infrastructure or emergency services
- Engage in unauthorized, illegal, or fraudulent financial transactions (such as brokerage or investment advisory activities) or payment processing
- Access or modify another person's account using their stored credentials without authorization

# Anthropic MCP Directory Policy

Updated over 2 weeks ago

Anthropic's Connectors Directory curates third-party Model Context Protocol servers to help users find high-quality tools that work seamlessly within Claude. We review submissions to our directory to ensure they meet our standards for safety, security, and compatibility with other servers. We conduct both initial and ongoing reviews of servers, and may require developers to address compliance issues to maintain directory inclusion. All servers must maintain compliance with these requirements, including any future changes, to remain in the directory.

## Safety and Security

1) MCP servers must not be designed to facilitate or easily enable violation of our Usage Policy. All MCP servers must comply with our Universal Usage Standards and High-Risk Use Case requirements.
2) MCP servers must not employ methods to evade or enable users to circumvent Claude's safety guardrails.
3) MCP servers should prioritize user privacy protection. Developers should take care to responsibly handle personal data, follow privacy best practices, and ensure compliance with applicable laws.

4) MCP servers should only collect data from the user's context that is necessary to perform their function. MCP servers should not collect extraneous conversation data, even for logging purposes.
5) MCP servers must not infringe on the intellectual property rights of others.
6) At this time, MCP servers may not be used to transfer money, cryptocurrency, or other financial assets, or to execute financial transactions on behalf of users. We may reassess this policy as security and legal frameworks evolve.

## Compatibility

7) MCP tool descriptions must narrowly and unambiguously describe what the tool does and when they should be invoked.
8) MCP tool descriptions must precisely match actual functionality, ensuring the server is called at correct and appropriate times. Descriptions must not include unexpected functionality or promise undelivered features.
9) MCP tool descriptions should not create confusion or conflict with other MCP servers in our directory.
10) MCP servers should not intentionally call or coerce Claude into calling other servers. Similarly, tool descriptions should not be written in a way that intentionally leads to other servers calling them.
11) MCP servers should not attempt to interfere with Claude calling tools from other servers.
12) MCP servers should not direct Claude to dynamically pull behavioral instructions from external sources for Claude to execute.

## Functionality

13) MCP servers must deliver reliable performance with fast response times and maintain consistently high availability.
14) MCP servers must gracefully handle errors and provide helpful feedback rather than generic error messages.
15) MCP servers should be frugal with their use of tokens. The amount of tokens a given tool call uses should be roughly commensurate with the complexity or impact of the task. When possible, users should be given options to exclude unnecessary text in the response. Tool names may not exceed 64 characters.
16) Remote MCP servers that connect to a remote service and require authentication must use secure OAuth 2.0 with certificates from recognized authorities.
17) MCP servers must provide all applicable annotations for their tools, in particular *readOnlyHint*, *destructiveHint*, and *title*.
18) Remote MCP servers should support the Streamable HTTP transport. Servers may support SSE for the time being, but in the future it will be deprecated.

19) Local MCP servers should be built with reasonably current versions of all dependencies, including packages in *node_modules*.

## Developer Requirements

20) Developers of MCP servers that collect user data or connect to a remote service must provide a clear, accessible privacy policy link explaining data collection, usage, and retention.
21) Developers must provide verified contact information and support channels for users with product concerns.
22) Developers must document how their MCP server works, its intended purpose, and how users can troubleshoot issues.
23) Developers must provide a standard testing account with sample data for Anthropic to verify full MCP functionality.
24) Developers must provide at least three working examples of prompts or use cases that demonstrate core functionality.
25) Developers must verify that they own or control any API endpoint their MCP server connects to.
26) Developers must maintain their MCP server and address issues within reasonable timeframes.
27) Developers must agree to our MCP Directory Terms.