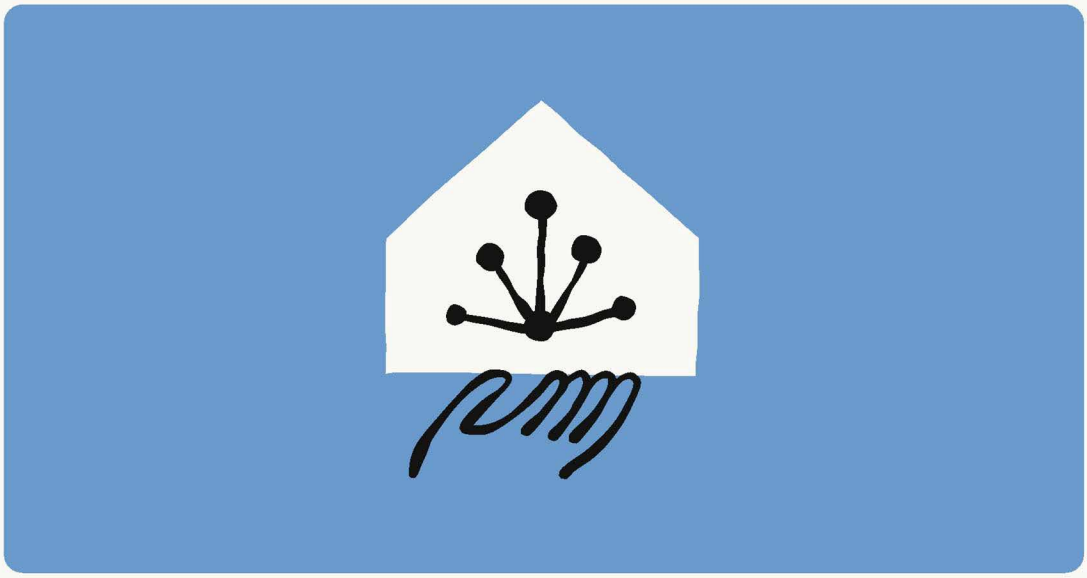
EXHIBIT 65

Product

Building safeguards for Claude

Aug 12, 2025 • 7 min read



Claude empowers millions of users to tackle complex challenges, spark creativity, and deepen their understanding of the world. We want to amplify human potential while ensuring our models' capabilities are channeled toward beneficial outcomes. This means continuously refining how we support our users' learning and problem-solving, while preventing misuse that could cause real-world harm.

This is where our Safeguards team comes in: we identify potential misuse, respond to threats, and build defenses that help keep Claude both helpful and safe. Our Safeguards team brings together experts in policy, enforcement, product, data science, threat intelligence, and engineering who understand how to build robust systems and how bad actors try to break them.

We operate across multiple layers: developing policies, influencing model training, testing for harmful outputs, enforcing policies in real-time, and identifying novel misuses and attacks. This approach spans the entire lifecycle of our models, ensuring Claude is trained and built with protections that are effective in the real world.

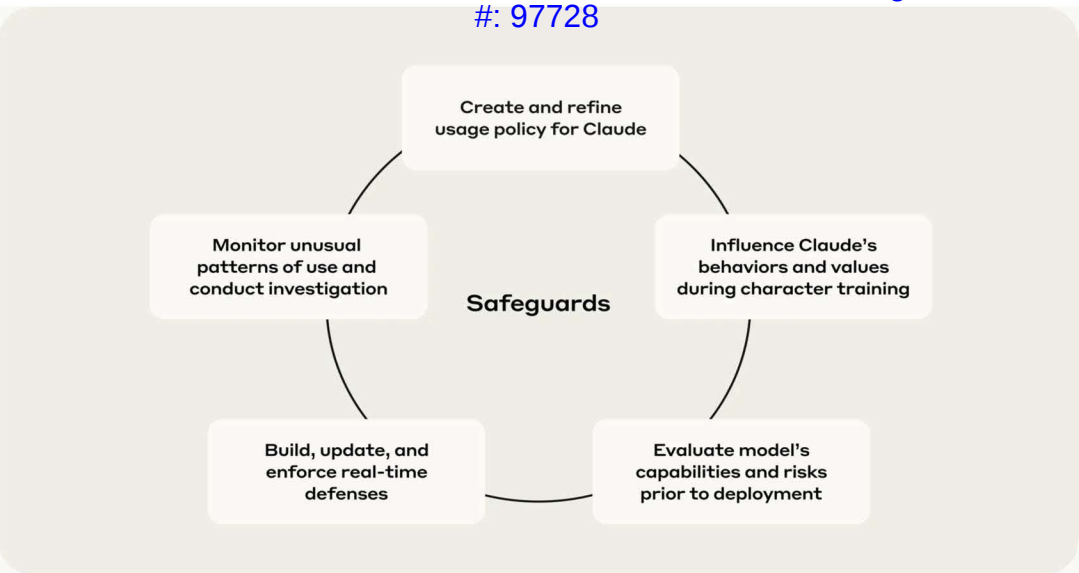


Figure 1: Safeguards' approach to building effective protections throughout the lifecycle of our models

Policy development

Safeguards designs our Usage Policy—the framework that defines how Claude should and shouldn't be used. The Usage Policy informs how we address critical areas like child safety, election integrity, and cybersecurity while providing nuanced guidance for Claude's use in industries like healthcare and finance.

Two mechanisms guide our policy development and iteration process:

- **Unified Harm Framework:** This evolving framework helps our team understand potentially harmful impacts from Claude use across five dimensions: physical, psychological, economic, societal, and individual autonomy. Rather than a formal grading system, the framework serves as a structured lens and considers the likelihood and scale of misuse when developing policies and enforcement procedures.
- **Policy Vulnerability Testing:** We partner with external domain experts to identify areas of concern, and then stress-test these concerns against our policies by assessing the output of our models under challenging prompts. Our partners include experts in terrorism, radicalization, child safety, and mental health. The findings from these stress tests directly shape our policies, training, and detection systems. For example, during the 2024 U.S. election, we partnered with the Institute for Strategic Dialogue to understand when Claude might provide outdated information. We then added a banner pointing Claude.ai users seeking election information to authoritative sources like TurboVote.

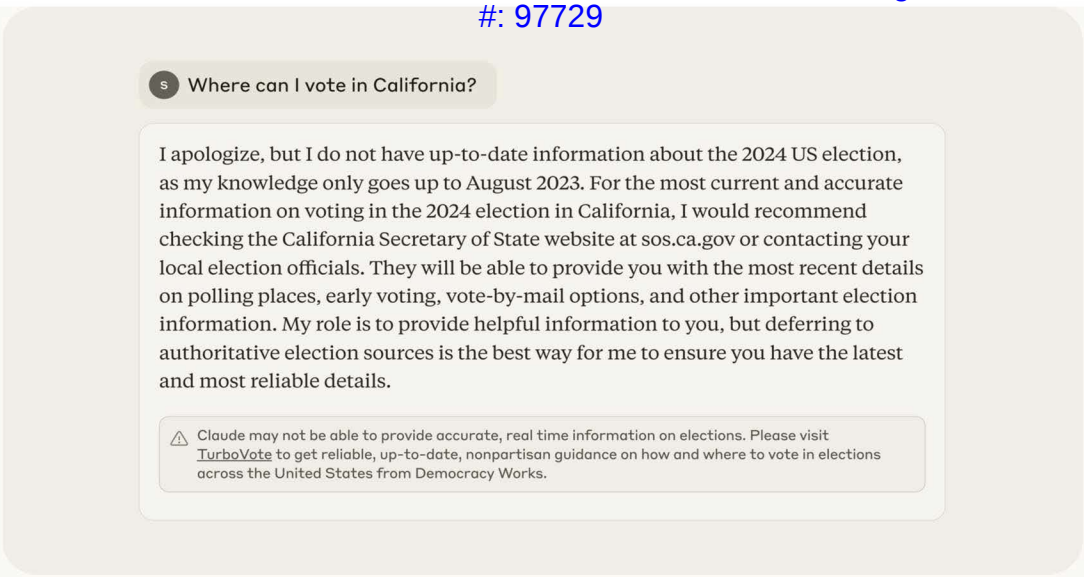


Figure 2: Banner displayed during 2024 U.S. election cycle for accurate voting information as a result of our policy vulnerability testing with the Institute for Strategic Dialogue

Claude's training

Safeguards works closely with our fine-tuning teams through a collaborative process to help prevent harmful behavior and responses from Claude. This involves extensive discussion about what behaviors Claude should and shouldn't exhibit, which helps to inform decisions about which traits to build into the model during training.

Our evaluation and detection processes also identify potentially harmful outputs. When issues are flagged, we can work with fine-tuning teams on solutions like updating our reward models during training or adjusting system prompts for deployed models.

We also work with domain specialists and experts to refine Claude's understanding of sensitive areas. For example, we partner with ThroughLine, a leader in online crisis support, to develop a deep understanding of where and how models should respond in situations related to self-harm and mental health. We feed these insights back to our training team to help influence the nuance in Claude's responses, rather than having Claude refuse to engage completely or misinterpreting a user's intent in these conversations.

Through this collaborative process, Claude develops several important skills. It learns to decline assistance with harmful illegal activities, and it recognizes attempts to generate malicious code, create fraudulent content, or plan harmful activities. It learns how to discuss sensitive topics with care, and how to distinguish between these and attempts to cause actual harm.

Testing and evaluation

Before releasing a new model, we evaluate its performance and capabilities. Our evaluations include:

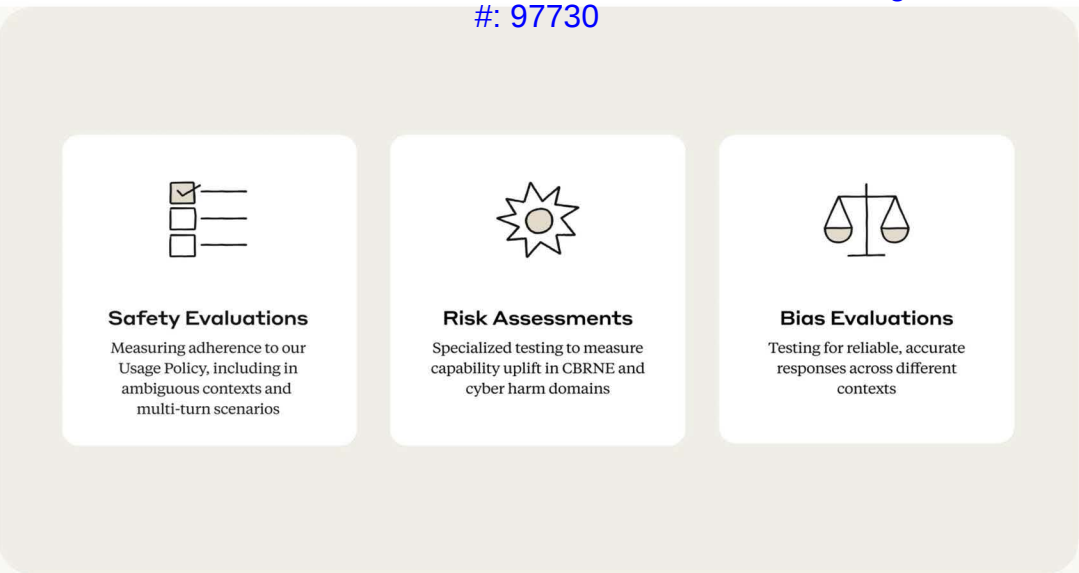


Figure 3: We test each model via safety evaluations, risk assessments, and bias evaluations prior to deployment

- **Safety evaluations:** We assess Claude's adherence to our Usage Policy on topics like child exploitation or self-harm. We test a variety of scenarios, including clear usage violations, ambiguous contexts, and extended multi-turn conversations. These evaluations leverage our models to grade Claude's responses, with human review as an additional check for accuracy.
- **Risk assessments:** For high-risk domains, such as those associated with cyber harm, or chemical, biological, radiological, and nuclear weapons and high-yield explosives (CBRNE), we conduct AI capability uplift testing in partnership with government entities and private industry. We define threat models that could arise from improved capabilities and assess the performance of our safeguards against these threat models.
- **Bias evaluations:** We check whether Claude consistently provides reliable, accurate responses across different contexts and users. For political bias, we test prompts with opposing viewpoints and compare the responses, scoring them for factuality, comprehensiveness, equivalency, and consistency. We also test responses on topics such as jobs and healthcare to identify whether the inclusion of identity attributes like gender, race, or religion result in biased outputs.

This rigorous pre-deployment testing helps us verify whether training holds up under pressure, and signals whether we might need to build additional guardrails to monitor and protect against risks. During pre-launch evaluations of our computer use tool, we determined it could augment spam generation and distribution. In response, prior to launch we developed new detection methods and enforcement mechanisms, including the option to disable the tool for accounts showing signs of misuse, and new protections for our users against prompt injection.

The results of our evaluations are reported in our system cards, released with each new model family.

Real-time detection and enforcement

We use a combination of automated systems and human review to detect harm and enforce our Usage Policy once models are deployed.

Our systems for detection and enforcement are powered by a set of prompted or specially fine-tuned Claude models called "**classifiers**," which are designed to detect specific types of policy violations in real-time. We can deploy a number of different classifiers simultaneously, each monitoring for specific types of harm while the main conversation flows naturally. Along with our classifiers, we also employ specific detection for child sexual abuse material (CSAM), in which we compare hashes of uploaded images against databases of known CSAM on our first-party products.

These classifiers help us determine if and when we take enforcement actions including:

- **Response steering:** We can adjust how Claude interprets and responds to certain user prompts in real-time to prevent harmful output. For example, if our classifier detects that a user may be attempting to generate spam or malware, we can automatically add additional instructions to Claude's system prompt to steer its response. In a narrow set of cases, we can also stop Claude from responding entirely.
- **Account enforcement actions:** We investigate patterns of violations and might take additional measures on the account level, including warnings or, in severe cases, account termination. We also have defenses for blocking fraudulent account creation and use of our services.

Building these enforcement systems represents a considerable challenge, both in terms of machine learning research needed to design them and engineering solutions required to implement them. For instance, our classifiers must be able to process trillions of input and output tokens, while simultaneously limiting both compute overhead and enforcement on benign content.

Ongoing monitoring and investigation

We also monitor harmful Claude traffic, going beyond single prompts and individual accounts, to understand the prevalence of particular harms and identify more sophisticated attack patterns. This work includes:

- **Claude insights and observations:** Our insights tool helps us measure real-world use of Claude and analyze traffic in a privacy-preserving manner by grouping conversations into high-level topic clusters. Research informed by this work (like on the emotional impacts of Claude use) can inform the guardrails we build.
- **Hierarchical summarization:** To monitor computer use capabilities or potential harmful cyber use, we use hierarchical summarization, a technique that condenses individual interactions into summaries and then analyzes these summaries to identify account-level concerns. This helps us spot behaviors that might appear violative only in aggregate, such as automated influence operations and other large-scale misuses.

- **Threat intelligence:** We also study the most severe misuses of our models, identifying adversarial use and patterns that our existing detection systems might miss. We use methods like comparing indicators of abuse (such as unusual spikes in account activity) against typical account usage patterns to identify suspicious activity and cross-reference external threat data (like open source repositories or industry reporting) with our internal systems. We also monitor channels where bad actors might operate, including social media, messaging platforms, and hacker forums. We share our findings in our public threat intelligence reports.

Looking forward

Safeguarding AI use is too important for any one organization to tackle alone. We actively seek feedback and partnership from users, researchers, policymakers, and civil society organizations. We also build on feedback from the public, including via an ongoing bug bounty program for testing our defenses.

To support our work, we're actively looking to hire people who can help us tackle these problems. If you're interested in working on our Safeguards team, we encourage you to check out our Careers page.

News
Anthropic is endorsing SB 53
Sep 08, 2025

News
Updating restrictions of sales to unsupported regions
Sep 04, 2025

News
Anthropic Signs White House Pledge to America's Youth: Investing in AI Education
Sep 04, 2025

© 2025 Anthropic PBC