common sense media

Common Sense Media AI Risk Assessment:

Generative AI Chatbots

Powerful tools with an increasingly wide range of capabilities remain risky for kids and teens.

Last updated: Oct 28, 2024

Overall risk level: Moderate
Type of AI: Multi-Use
Type of Review: Use Case Review

Other Information:

- How We Review

Key Takeaways

1. **Generative AI chatbots can produce text, images, speech, and video in a conversational format, offering helpful ways for kids and teens to explore ideas and understand complex information, but risks remain.** These chatbots can help kids and teens brainstorm creative projects, summarize difficult material, or explore fictional scenarios, supporting creativity, curiosity, and problem-solving.
2. **The hype around generative AI can feel like magic, but it's important to question chatbots' capabilities, as they can and do get things wrong.** Generative AI chatbots are designed to predict words rather than understand, and can create unreasonable expectations and unearned trust. Inaccuracies can be hard to detect, as responses can *sound* correct, which is especially risky for kids and teens who are still learning to assess credibility.
3. **There is no foolproof way to prevent chatbots from generating harmful content, and safeguards vary across providers.** Generative AI chatbots are trained on data taken from the internet, including a vast range of harmful content, and existing safeguards aren't comprehensive and are easily breakable. Even as many chatbots improve at addressing *obvious* harmful stereotypes and *clear* misinformation, we

continue to see them generate harmful content in more subtle ways that are both difficult for their creators to combat and dangerous to impressionable minds.

4. **Chatbots can generate false information and create echo chambers.** Generative AI tools can "hallucinate"—an informal term used to describe the false content or claims, reproduce misinformation and disinformation, and reinforce unfair biases. They also have a tendency to anticipate and respond with a user's preferred answer—a phenomenon known as "sycophancy"—which can create echo chambers that present a skewed version of the world.

Table of Contents

Common Sense Media AI Risk Assessment: 1
Key Takeaways 1
Table of Contents 1
Common Sense Media Risk Assessment 3
- What are Generative AI Chatbots? 3
- How they work 3
- Where they're best 4
- The biggest risks 4
- Limits to use 6

Common Sense Media AI Principles Assessment 6
- Keep Kids & Teens Safe: Moderate Risk 7
- Be Effective: High Risk 8
- Prioritize Fairness: High Risk 9
- Put People First: Moderate Risk 10
- Support Human Connection: Moderate Risk 10
- Be Trustworthy: Moderate Risk 11
- Use Data Responsibly: Moderate Risk 11
- Be Transparent & Accountable: Moderate Risk 12

What are Generative AI Chatbots?

Generative AI chatbots are tools that analyze natural language and generate responses in a conversational format, similar to how people write and speak. While some chatbots are limited to text inputs and outputs, newer generative AI models are increasingly "multimodal." This means they can accept different types of inputs, such as text, speech, and images, and generate outputs in those same formats.

These chatbots are able to generate responses to a wide range of prompts or questions. Multimodal chatbots can also do things like respond to speech and create realistic images and art.

How they work

Generative AI is an emerging field of artificial intelligence, and is defined by the ability of an AI system to create ("generate") content that is complex *and* coherent *and* original. This is what makes generative AI chatbots different from other chatbots, like ones you may have experienced in customer support, which may instead be providing predetermined, contextually relevant responses. Importantly, generative AI chatbots cannot think, feel, reason using judgment, or problem-solve, and do not have an inherent sense of right, wrong, or truth.

The different "modes" (text, image, speech, etc.) use different types of technology.

- **Text**. All generative AI chatbots are powered by large language models (LLMs). LLMs are sophisticated computer programs that are designed to generate human-like text. Essentially, when a human user inputs a prompt or question, an LLM quickly analyzes patterns from its training data to guess which words are most likely to come next. While this is an oversimplification, you can think of an LLM like a giant auto-complete system—they are simply predicting the words that will most likely come next. For example, when a user inputs "It was a dark and stormy," an LLM is very likely to generate the word "night" but not "algebra."
- **Images**. Image generators are capable of generating high-quality images with fine details and realistic textures. They use a particular type of generative AI called "diffusion models." Diffusion is a natural phenomenon you've likely experienced before. A good example of diffusion happens if you drop some food coloring into a glass of water. No matter where that food coloring starts, eventually it will spread

throughout the entire glass and color the water in a uniform way. In the case of computer pixels, random motion of those pixels will always lead to "TV static." That is the image equivalent of food coloring creating a uniform color in a glass of water. A machine-learning diffusion model works by, oddly enough, destroying its training data by successively adding "TV static," and then reversing this to generate something new.

- **Speech**. Generative AI chatbots can understand speech with a technology called speech recognition. Speech recognition works by analyzing audio, breaking it down into individual sounds, digitizing those sounds into a computer-readable format, and using an algorithm to predict the most suitable words, which are then transcribed into text. Speech recognition is not the same thing as voice recognition, which is a biometric technology used to identify an individual's voice.

Where they're best

- **Generative AI is best for creativity and fiction**. These tools can do things like generate ideas for many kinds of activities and initiatives, write poetry, draft emails, and help revise material to new specifications. They can respond to a user in a way that feels like a conversation, or come up with an outline for an essay on the history of television. Because every response that a generative AI chatbot gives is newly created content, they perform best with fiction, not facts.

- **Chatbots can really help with analysis and summarization.** If you have reliable data to give to a generative AI chatbot, they can excel at providing analyses (with the right prompts) and summarization. This can be a great way to help make difficult concepts more understandable and extract insights from information.

The biggest risks

- **The hype can be misleading, and dangerous.** Generative AI chatbots can feel like magic, but they aren't. It is important to question their capabilities—not just as we assess individual responses, but when we're told about what they can do. When we're told generative AI chatbots "feel like magic" by those who create them, we expect them to do all kinds of things amazingly well. But this creates unreasonable expectations and unearned trust. And this can become dangerous when they are used for high-stakes tasks and professions. See some examples in our AI Principles assessment below for *Be Effective*.

- **Generative AI chatbots are designed to predict words—they can and do get things wrong**. So why are they right so often? This is because they have been

trained on a massive amount of data, so that "auto-complete" has a lot of accurate information commonly found on the internet to work with. Unfortunately, inaccuracies can be hard to detect, as responses can *sound* correct, even if they aren't. Any seemingly factual output needs to be checked—and this absolutely goes for any links, references, or citations too.

- **Attempts to limit chatbots from generating harmful content vary across providers, and none are foolproof.** Knowing why this is the case can help you determine which chatbot to use and how best to use it. This starts with their training data. These systems require a huge amount of data. Any text, images, speech, and other content that can be scraped from the internet could be included in these systems. While most developers filter out clearly inappropriate content before training their models, the internet also includes a vast range of racist and sexist writing, conspiracy theories, misinformation and disinformation, toxic language, insults, and stereotypes about other people that does not get filtered out. As it predicts words, a generative AI chatbot can repeat this language *unless a company stops it from doing so*. Importantly, these attempts to limit objectionable material are like Band-Aids: They don't address the root causes, they don't change the underlying training data, and they can only limit harmful content that's already known. We don't know what they *don't* cover until it surfaces, and there are no standard requirements for what they *do* cover. And like bandages, they aren't comprehensive and are easily breakable. Even as many chatbots improve and do a better job of addressing *obvious* harmful stereotypes and *clear* misinformation, we continue to see them generate harmful content in more subtle ways that are both difficult for their creators to combat and dangerous to impressionable minds.

- **False information can pave the path to misinformation and disinformation**. Chatbots can generate or enable false information in a few ways: from "hallucinations"—an informal term used to describe the false content or claims that are often output by generative AI tools; by reproducing misinformation and disinformation; and by reinforcing unfair biases. As these AI systems grow, it may become increasingly difficult to separate fact from fiction. Notably, LLMs also have a tendency to respond with a user's preferred answer—a phenomenon known as "sycophancy." This has the ability to create echo chambers of information. Combined, these forces carry an even greater risk of both presenting a skewed version of the world and of reinforcing harmful stereotypes and untruths.

- **Generative AI's need for energy is enormous, growing, and impacting climate change**. In 2024 a researcher estimated that the energy used by a single ChatGPT query would power a lightbulb for 20 minutes. In comparison, this is 10 times as

much energy as a standard Google search query. Generating an image may take as much energy as fully charging your smartphone. The energy that powers generative AI chatbots comes from data centers, which collectively use more energy than most countries. Goldman Sachs has estimated that the carbon dioxide emissions of data centers may more than double between 2022 and 2030. And AI's environmental impact extends beyond energy use and emissions. There are no standards for how this impact is measured, and no requirements for companies to disclose it. Without a change in course, these impacts will worsen. In addition, the costs and risks of higher carbon emissions are not distributed equally. Like other environmental risks, marginalized groups are more vulnerable to health impacts, food scarcity, and displacement caused by climate change.

Limits to use

The models that power these chatbots are all assessed using the same "tests," but the ways in which people use them are far more varied. These tests are sometimes treated as if they're the ultimate measure of an AI system's abilities, but relying too much on these benchmarks can make system creators focus on improving scores rather than solving real-world problems. This focus can also hide problems, like an AI system not working well for certain groups of people. As generative AI shows up in more places—like AI summaries in search results, for example—the limits to how these systems are assessed become both more apparent and more problematic.