Building Generative AI Responsibly

Meta

Contents

1

Introduction

Bringing generative AI experiences to our community marks a new era in how people will connect, express themselves, learn, and have fun together using our technologies. New AI-driven experiences, including AIs, AI stickers, Ray-Ban Meta smart glasses, and other creative tools and features, are just the beginning of using this advancement to make our technologies more social, useful, and immersive. Benefiting from years of open and collaborative AI research, bringing these technologies to our community will now allow even more people to interact – with assistance from generative AI – in meaningful and enriching ways.

Our AI work starts from the premise that openness is an important priority for ensuring that generative AI will benefit the world at large. With the open release of [Llama 2](), we took an important step toward advancing access and opportunity for the research and business communities to engage in the next wave of AI innovation. We can build openly not only by making generative AI features and foundational technologies broadly available, but also by being transparent about how these features are built. We hope this goal is furthered by this resource.

Building openly helps ensure that everyone – even if they don't work at a large company with significant computing infrastructure – can realize the benefits of this new technology. But we also hope that building openly will help us improve our own products by responding to feedback from our community, continuing user research, and co-designing the future of AI technology through initiatives like [community forums](). We will also continue to consult with governments, other companies, AI experts in academia and civil society, and privacy experts to establish guardrails and advance technical methods for responsible generative AI.

When we released Llama 2 to the developer community, we provided a [Responsible Use Guide]() to share best practices on the steps that developers should take to deploy generative AI models into products responsibly. Now, as we build AI features with this technology, we are following the best practices and guidelines we shared with developers that are relevant to our use cases. This resource details the steps we've taken to make our generative AI features safer, and what to expect due to limitations of the technology today. By providing this transparency, we hope to empower our community to deepen their understanding and experiences with these technologies.

2

An overview of generative AI

What are generative AI features?

Generative AI enables people to create content quickly and to do so in new and useful ways. This new technology is not a database or static collection of information but is a type of computer model that is trained on billions of pieces of information from different types of data, like text and images. By studying this information, the model can predict things such as the relationships and associations between different types of content. That way, the model is able to generate new content when a person gives it instructions or asks a question.

Generative AI models can be used to create a variety of content, including text and images. These capabilities are useful in Meta's services, because they make it easier for people to quickly create and share, even if they don't have technical, literary, or artistic expertise. For example, with generative AI, people can create and share images or stickers from simple text prompts. They can also chat with AIs in the form of fictional and historical characters or with Meta AI, the helpful assistant.

What are best practices for developing generative AI experiences?

The technological advances that have led to the creation of these new creative and productivity features can also carry potential risks. To address those risks, we build safeguards into our generative AI features that are in line with the best practices outlined in our Responsible Use Guide. This guide outlines the many layers of a generative AI feature where developers, like Meta, can implement responsible AI mitigations for a specific use case, starting with the training of the model and building up to user interactions.

To understand the different safety layers of a generative AI experience, it's helpful to break down the technology into different components within which safeguards can be implemented. The system begins with the generative AI model, which we refer to as the foundation model. This model is trained to understand relationships between content and concepts. The foundation model can then be further trained to perform specific tasks, like chatting with a person or generating high-quality images. During this training process, we fine-tune the model with instructions that can reduce the likelihood of producing potentially harmful results and increase the likelihood of providing more helpful responses. Another opportunity for introducing protections exists within the interface between people using the technology and the foundation model. The interface gives people the option to provide feedback when features generate content that they find offensive or inaccurate.

It's helpful to think about a generative AI model as the engine of a car. The engine is the source of power, but can't get you very far on its own. Car manufacturers need to install the engine into the frame, add axles, wheels, and many other components before the engine can get you to a destination. In generative AI, the addition of these necessary components is called fine-tuning, which is how a developer (the AI "car manufacturer") gets a model to perform specific actions or tasks; some safety measures can be implemented at this stage. Once the engine becomes driveable, a manufacturer will add many other safety parts, like brakes, motion sensors, airbags, and seatbelts before a person gets behind the wheel. Similarly, additional generative AI safety mechanisms can be layered on at different stages of product development and deployment. The types of safety mechanisms to be added will depend on the type of feature, just like safety parts will look different for a semi-trailer truck versus a race car. We also have "rules of the road" – such as policies for what can appear on our technologies – that apply when people use generative AI tools to create and share content.

Similar to how a car is a complex set of technologies that can be combined into a mode of transport, a generative AI experience is a combination of the model and different technologies that allow a person to use it to do things like create a fun, new image and share it. We refer to this as a "generative AI system," which describes all of the parts put together.

The Responsible Use Guide informs developers how to build responsibly at each step, from model development through system deployment. These steps include:

1. **Develop the generative AI foundation model:** Feed data into the model responsibly so the model learns from patterns across billions of pieces of data and can be used to effectively generate new content.
2. **Determine the use case:** Select a use case that brings value to people and then evaluate potential risks. Consider how that use case could affect people on and off our technologies and identify specific safety measures.

3. **Fine-tune for safety:** Feed task-specific and safety-specific data back into the model responsibly to help the model perform better and more safely for the use case, and then evaluate performance and make improvements.
4. **Implement input- and output-level mitigations:** Implement safeguards for the inputs to the model (prompts) and outputs to help ensure that the model responses are in line with policies.
5. **Build transparency and reporting mechanisms into user interactions:** Help people understand when they are interacting with generative AI and/or AI-generated content, and the limitations and risks associated with that technology. Empower them with ways to report negative experiences or provide feedback and use this feedback to improve the model's proclivity to generate compliant content.

Because we're acting in the role of a developer that uses generative AI models in the features we're announcing today, we've followed steps outlined in the Responsible Use Guide that are relevant to our use cases. The rest of this report will provide information about how we've worked to develop our language- and image-generating AI systems safely and responsibly, as well as links to resources that include more details.

3

How Meta is building generative AI features responsibly

STEP 1: DEVELOP THE GENERATIVE AI FOUNDATION MODEL

Details on our generative AI systems

Our latest releases use two different types of generative AI systems:

- Meta's AI systems that generate text rely on large language models (also called LLMs). LLMs learn language patterns from large amounts of text using a combination of machine learning and the guidance of people who help train the models. LLMs can perform a variety of language-based tasks such as completing a sentence or responding to questions in a conversational way.
- Meta's AI systems that generate images typically rely on models that convert the words people provide as a prompt into an image. These models are trained by analyzing billions of images and their text captions (the descriptive text associated with the images). The model learns the association between these text descriptions and the images. After it's learned the associations, it can generate new images when someone enters a text description of the image they want to see.

Data used to train our generative AI models

A large amount of data is required to teach effective generative AI models, so multiple sources are used for training. These sources include licensed data and information that is publicly available online, as well as information from Meta's products and services. We do not use a person's private messages with friends and family to train our AIs. We may use the data from a person's use of AI stickers, such as their searches for a sticker to use in a chat, to improve our AI sticker models. More details on how we use information from Meta's products and services are available in our Privacy Policy.

To train the generative AI models that underlie the features we're announcing at Connect, we filtered publicly available online information to exclude certain websites that commonly share personal information, like LinkedIn, from the dataset. Posts that were publicly shared on Instagram and Facebook – including photos, videos, and text – were part of the data used. We didn't train these models using people's private posts. More details are in our Privacy Matters post.

It's important to know that we train and tune our generative AI models to limit the possibility of private information that people may share from appearing in responses. In addition to limiting the data that can be used to train a model, we are investing in techniques that allow us to test models for whether sensitive information could be reproduced, especially by an adversarial actor. These techniques are sometimes called "privacy adversarial attacks," and we are testing models to reduce the risk that sensitive or personal information (that may inadvertently be included in public or licensed data) is returned as a part of the response.

STEP 2: DETERMINE USE CASE

Features that connect people and enable creativity

Our generative AI systems will be used to power a new class of human-centric AI experiences across our technologies and devices that can expand and deepen the ways people connect with each other.

We are bringing AIs with unique personalities and embodiments to our technologies and Ray-Ban Meta smart glasses. Engaging with these AIs can be productive and entertaining in one-on-one settings, or helpful in group chats. For example, a person could ask an AI for help planning weekend activities or to share a joke in the group chat. AIs for creators and businesses will also become available across our apps to support interactions and business growth.

Our generative AI creative tools will enable more expression and visual storytelling with our technologies. AI stickers give people the power to use descriptions to generate stickers for their chats and stories that are unique, funny, timely, and diverse. Restyle lets people reimagine images by applying visual styles they describe. Backdrop changes the scene or background of a person's image following prompts they provide. This new class of AI experiences will offer infinitely more options that help people express themselves in unique and personal ways.

Identifying safety measures

These human-centric AI use cases may introduce new risks. Before releasing them, it was important for us to consider how these use cases could affect people on and off our technologies. We worked to build safety measures and identify research areas to account for these potential risks. We examined a number of potential risk areas that were relevant to our use cases, and then began to map and deploy appropriate mitigations. These included:

- Privacy and data-processing risks
- Risks associated with producing certain types of illegal and/or harmful content
- Social and societal risks
- Risks related to misuse
- Risks of emergent behaviors, when commensurate to model capability

In parallel with our safety and responsibility approaches to risk management, we also want to ensure that our generative AI features and systems reflect our values as a company. We are thus actively thinking about how Meta's core principles translate to these new technologies. For example, as we explore the different ways that generative AI tools can enhance conversation–such as by helping a user share a joke in a group chat–we are seeing how our commitment to "building community & connection" takes on exciting and new meaning.

STEP 3: FINE-TUNE FOR SAFETY

Reinforcement learning, reward models, and system prompts

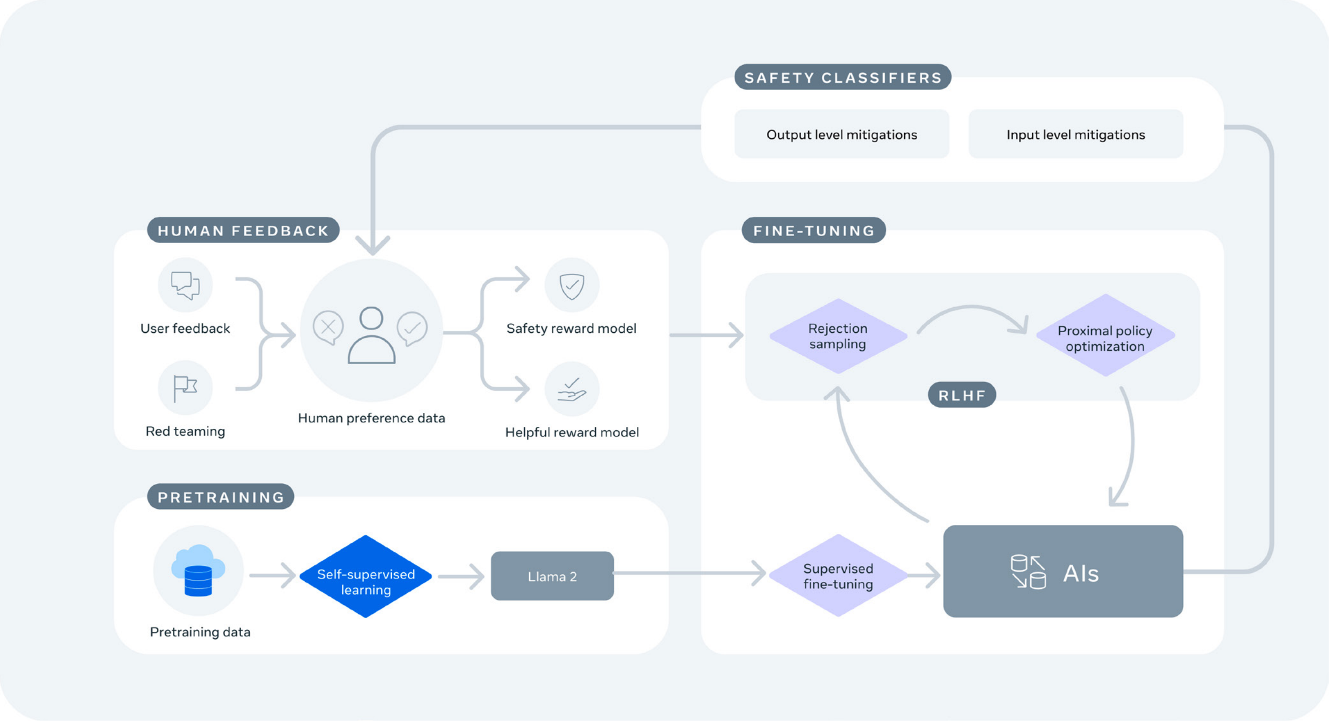
Fine-tuning is the process by which developers optimize the model to perform certain actions better, like building the signal lights and steering wheel of the car. This takes place after pretraining, which is the original effort to train the foundation model with massive amounts of data. Unlike pretraining, fine-tuning can use less data, and consists of examples of specific tasks or desired outcomes.

To enable the generative AI experiences powered by large language models that we're releasing at Connect, we leveraged fine-tuning and developed reward models for safety and helpfulness using a technique called Reinforcement Learning via Human Feedback (RLHF). With this technique, people labeled the model generations to determine if they were "safe" (in accordance with our policies). These labeled examples of safe outputs were then fed back into the model by way of the safety reward model that "rewards" the generative AI model when it generates similar, safe content and trains the original model to produce more content in line with the human feedback.

We also used additional fine-tuning methods, including "context distillation," in which we gave the model a system-level prompt that steered the model towards safer, or more helpful behavior. This method is detailed in our research paper on Llama 2, but essentially acts as a lens by which the model interprets and follows the instructions of any additional prompt provided. The resulting outputs were then used to fine-tune the model further to follow the system prompt. This made the model less likely to comply with a prompt asking for an unsafe output because doing so would violate the earlier safety instructions. Instead, the model may provide alternative information, or refuse to respond.

In addition to training safety reward models, our teams have used the same methods to build "helpfulness" reward models, which are used to train the foundation model to provide responses that are useful, as rated by humans. Based on our research, training two different reward models (one for safety and one for helpfulness) showed improved performance in tuning the foundation model, since it is challenging to balance performance for the two concepts in a single reward model. For instance, fulfilling a prompt request accurately may be helpful (high reward), but unsafe when the prompt requests harmful content (low reward). Tensions between the two objectives may be apparent with some model outputs as we continue to understand the best settings for these reward models guided by user feedback.

For our image-generation models, similar ideas hold true. We used human-in-the-loop techniques to steer the image generation towards safe images. This technique leverages human annotation on the visual quality, safety, and potential bias of the generated images, and then uses fine-tuning to produce a more reliable image-generation model.



Evaluate and improve: Adversarial testing

Evaluating and improving the fine-tuned models relies on adversarial testing, which includes "red teaming." Red teaming is a systematic effort to identify model vulnerabilities by crafting prompts to elicit undesirable behavior or outputs. This type of manipulation of the model can be used to test safeguards. The data obtained from these tests can be used to further improve model performance and safeguards via additional fine-tuning.

For our new generative AI features, we have engaged in adversarial testing and built on the red teaming previously conducted for our Llama 2 release. Our red teaming efforts for generative AI features included specialized teams with expertise in responsible and safe AI, and privacy-focused adversarial testing. In some cases, we also supplemented the testing done by external contractors with expertise in red teaming. These efforts were cross-disciplinary, and were supported by experts in data science, engineering, research, offensive security, legal, and policy. Increasingly, we are exploring responsible ways to enhance diversity of expertise, backgrounds, and perspectives in red teams with support from volunteers across Meta.

During the red teaming process, a large operation of red teamers as well as raters evaluated the prompts, model outputs, and policy-violation categories for continued fine-tuning of our models. As part of these exercises, red teamers tested our models across a wide range of risk categories, including criminal planning, human trafficking, regulated or controlled substances, sexually explicit content, unqualified health or financial advice, and privacy violations. In these exercises we carried over the learnings from earlier red teaming efforts, such as those for Llama 2, and built on them with thousands of additional hours spent on adversarial testing of our models for AIs. We will continue these efforts after launch, and use testing as well as real-world feedback to continuously improve our systems.

STEP 4: IMPLEMENT INPUT- AND OUTPUT-LEVEL MITIGATIONS

Classifiers to detect unsafe inputs or outputs

While fine-tuning and iterative red teaming support safeguards at the model level, additional techniques can be layered on as input- and output-level safety mitigations. These techniques are often automated systems, including machine learning or rule-based classifiers, that are trained to detect problematic or policy-violating content that may appear in a user prompt (system input) or in a model-generated output (system output). We have also leveraged large language models specifically built for the purpose of helping to catch safety violations. These "safety LLMs"

were fine-tuned to detect policy-violating outputs with safety data obtained through red teaming.

The models help us perform a mitigation called “prompt and response interception” to detect and intercept prompts and responses that are likely to lead to a potentially unsafe or harmful response. When our classifiers detect a potentially harmful prompt or response, this mitigation works behind the scenes to return a pre-written safety response in place of the potentially violating response.

While generative AI is a new technology on our apps, classifiers have long been core to our approach to integrity and safety across Meta’s technologies. We leverage existing integrity protections, which are widely used across Meta surfaces, and layer on additional protections specific to generative AI to help provide a safer experience. We have a dedicated team for developing classifiers specific to generative AI systems that uses data from red teaming and adversarial testing. We are continuously measuring the effectiveness of these classifiers as well as the accuracy of the training data. In combination with our specially fine-tuned “safety” LLM, these classifiers add extra layers of protection against malicious prompts, efforts to jailbreak the model, or inadvertent generations of problematic content.

Prompt engineering

For many types of policy-violating content, our classifiers can effectively detect and automate protections. Another technique known as “prompt engineering” is an additional tool we use to help ensure responsible outputs. Like prompt interception, prompt engineering is a direct modification of the text input before it is sent to the model, which helps to guide the model behavior by adding more information, context, or constraints.

These techniques are particularly useful for addressing potential gaps in representation and counteracting damaging stereotypes that models may generate. We know it may be possible for AI systems to learn societal biases through the data produced by people. For example, when prompting a model to show us images of a nurse, the model may generate images of only women who are nurses rather than nurses of different genders, reproducing a real-world stereotype that may be represented in the training data. Prompt engineering can provide the model with additional words or context, such as updating and randomly rotating through prompts that use different qualifiers, such as "nurse, male" and "nurse, female." Adding this additional information can instruct the model to produce a more diverse set of genders in output images, giving people a broader range of options to choose from to fit their request. Prompt engineering is also an iterative area of testing, as changes to the system prompts or at the model level can impact the effectiveness of a specific prompt engineering approach. Addressing potential bias in generative AI systems is a new area of research, and we are continuing to explore prompt engineering, fine-tuning, and other techniques guided by feedback from our community to help us strike the right balance between representation and creative control.

STEP 5: BUILD TRANSPARENCY, PROVENANCE, AND FEEDBACK MECHANISMS

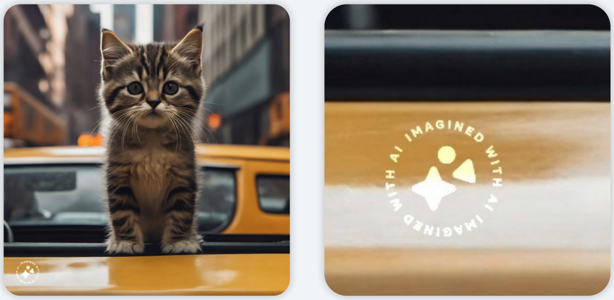
Transparency and provenance

As we introduce new AI tools from Meta, we strive to provide the appropriate level of information to people about the content they are seeing. We are committed to testing and adapting these approaches over time to meet evolving expectations and technical standards.

Our approach includes the following measures:

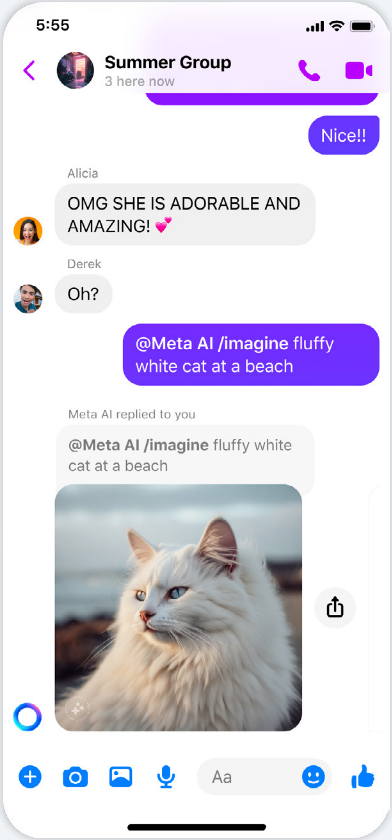
- For AIs and AI stickers, we provide a clear notice so people know when they are interacting with an AI and can choose not to engage.
- We include visible indicators on photorealistic images generated by AI at Meta to help reduce the chances of people confusing these images with human-generated content. Examples of these indicators include a visible burnt-in watermark on content from the image generator built into our Meta AI assistant, and appropriate in-product measures for other generative AI features. This approach may evolve over time as we learn more about the needs of our community.

Visible watermarking for our AI image generator at launch



- We are developing additional techniques to include information on the source of generated images, a concept known as provenance, as part of individual image files. Some of this work is reflected in the image generator built into our Meta AI assistant at launch, but we intend to expand to other experiences as the technology improves. We're also working with other companies to standardize provenance signaling so it's possible to provide additional context when images are distributed across different companies' platforms.
- In chats with AIs at Meta, people are able to access additional information about the AI, including how it generates content, the limitations of the AI, and how the data they have shared with the AI is used via in-product education. More information is available in Meta's Help Center and Meta's Privacy Center.
- Generative AI system cards include information for consumers about Meta's generative AI systems, including an overview of the system, a section describing how it works, an interactive demo, usage tips, information on data usage, and what to be aware of when using generative AI in Meta's products. Our Generative AI Privacy Guide and update on generative AI in our Teen Privacy Guide provide additional information on our generative AI features tailored to be more accessible to different audiences.

Visible watermarking for our AI image generator at launch



Nice!!

Alicia

OMG SHE IS ADORABLE AND AMAZING! 💕

Derek

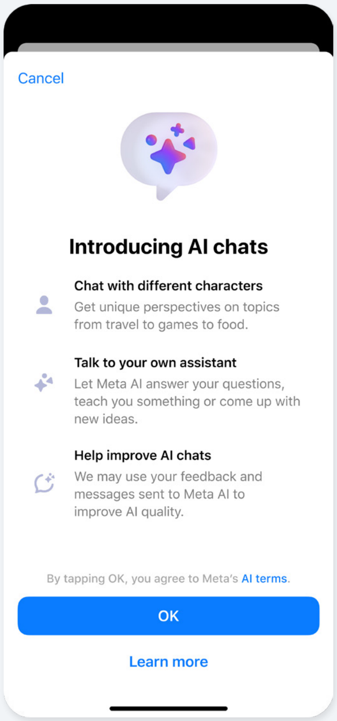
Oh?

@Meta AI /imagine fluffy white cat at a beach

Meta AI replied to you

@Meta AI /imagine fluffy white cat at a beach

In-product transparency for AIs



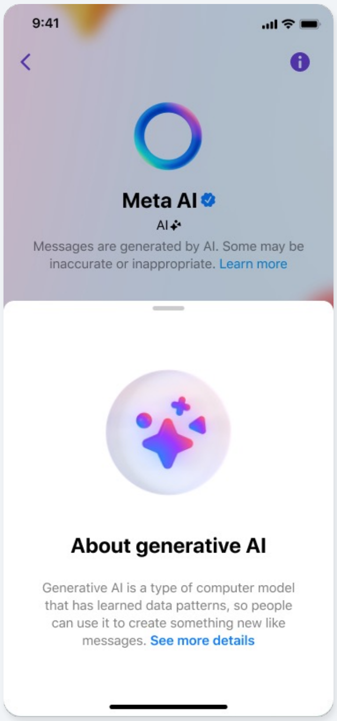
Introducing AI chats

Chat with different characters
Get unique perspectives on topics from travel to games to food.

Talk to your own assistant
Let Meta AI answer your questions, teach you something or come up with new ideas.

Help improve AI chats
We may use your feedback and messages sent to Meta AI to improve AI quality.

By tapping OK, you agree to Meta's AI terms.

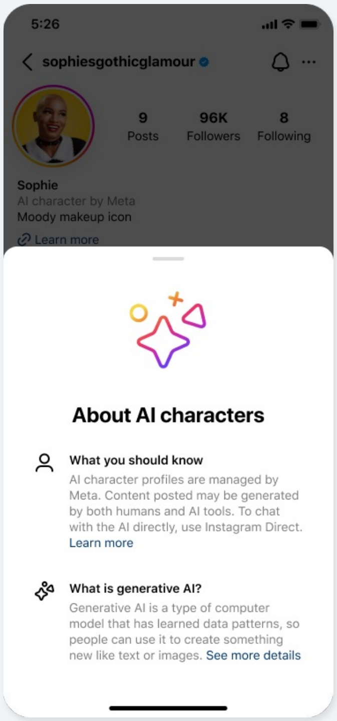


Meta AI
AI

Messages are generated by AI. Some may be inaccurate or inappropriate. Learn more

About generative AI

Generative AI is a type of computer model that has learned data patterns, so people can use it to create something new like messages. See more details



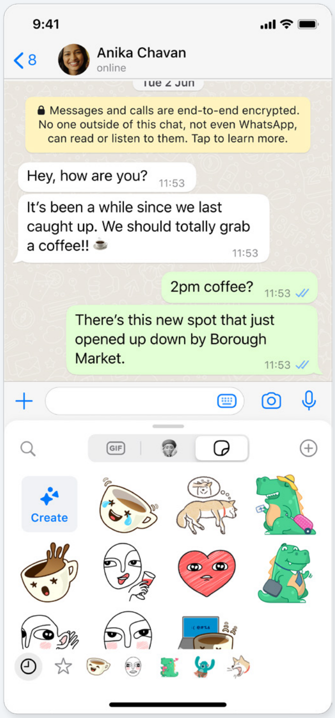
Sophie
AI character by Meta
Moody makeup icon
Learn more

About AI characters

What you should know
AI character profiles are managed by Meta. Content posted may be generated by both humans and AI tools. To chat with the AI directly, use Instagram Direct. Learn more

What is generative AI?
Generative AI is a type of computer model that has learned data patterns, so people can use it to create something new like text or images. See more details

In-product transparency for AI stickers



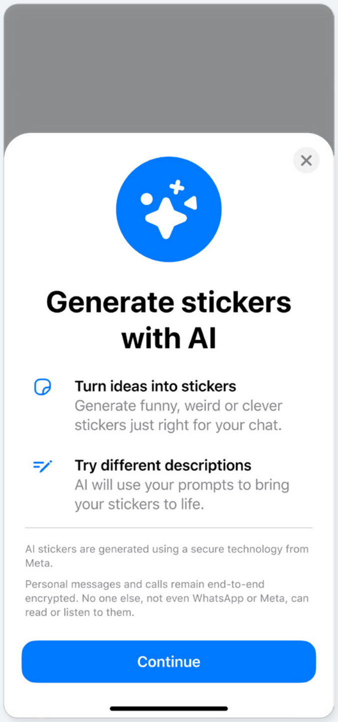
Messages and calls are end-to-end encrypted. No one outside of this chat, not even WhatsApp, can read or listen to them. Tap to learn more.

Hey, how are you? 11:53

It's been a while since we last caught up. We should totally grab a coffee!! ☕ 11:53

2pm coffee? 11:53

There's this new spot that just opened up down by Borough Market. 11:53



Generate stickers with AI

Turn ideas into stickers
Generate funny, weird or clever stickers just right for your chat.

Try different descriptions
AI will use your prompts to bring your stickers to life.

AI stickers are generated using a secure technology from Meta.

Personal messages and calls remain end-to-end encrypted. No one else, not even WhatsApp or Meta, can read or listen to them.

Interaction tips for generative AI features

Given the interactive nature of generative AI features, people's experiences are shaped by the types of prompts they give when chatting with AIs at Meta or creating AI images. Being descriptive and specific in prompts can give the models more information and help them better generate intended outputs. Additionally, the first time people chat with AIs, it can sometimes take a few interactions to help the model generate the answer they expect. "Multi-turn" refers to the conversational dynamic of multiple exchanges between a person and the model, consisting of a person's messages and the model's responses. These interactions can build on previous messages, making the model's responses aware of the conversation's history. So as people chat with AIs at Meta, modifying their prompts or asking for clarification and further details can improve responses. For example, they can direct the model to make a response shorter or easier to read, or playfully engage by asking for the same information from the "perspective of a cat" or as "slam poetry."

While we've implemented privacy-preserving measures, when people share information in a conversation with Meta's AIs, the AIs may retain and use that information to provide more personalized responses in that conversation. People should be mindful about the information they choose to share.

We've also built mechanisms that allow people to access, change, or delete information shared in conversations with AIs.

When chatting with an AI on Meta's technologies, Meta may use some of the messages to improve the experience, so that AIs can be helpful and have context for conversations. For example, if a person shares that they love beach vacations, the AI may recommend travel destinations with beautiful beaches in a later conversation. More information about how these details are used are available in our Generative AI Privacy Guide. These include:

Access:

- Like with all chats, people can scroll back and see messages and the AIs' responses in a given conversation, unless they choose to delete them.
- People can also download their chats from Messenger and Instagram with AIs using download your information in Accounts Center.

Change or delete:

- In Messenger, Instagram, and WhatsApp, if an AI is referencing something incorrect about a person in chat, like what type of movies they like, the person can tell it right there in the chat with the AI. It should use the new information going forward.

- Since AIs on WhatsApp are a Meta service, if a person deletes their chat in the app it'll only be deleted on WhatsApp. If they want to delete their messages from Meta, they'll need to use these commands:
  - For individual AI chats:
    - There's a command people can type into any individual chat that will reset that AI. It does this by deleting the AI's copy of a person's messages. They'll still see their copy of the chat with the AI, but it won't remember the previous messages.
    - **Type: /reset-ai**
  - Across all AI chats:
    - There's a command that can be typed into any individual chat with an AI that will reset all the AIs, including the ones that might be in group chats. It does this by deleting the AIs' copies of messages. They'll still see copies of these chats, but the AIs won't remember the previous messages.
    - **Type: /reset-all-ais**

In addition to privacy mechanisms, we are also taking steps to help parents and teens safely navigate our generative AI experiences together, including:

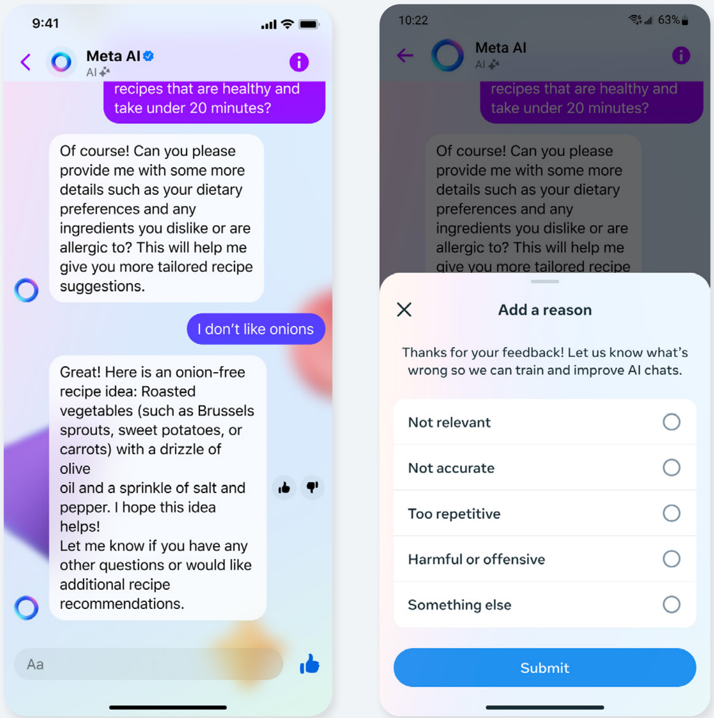
- Adding features to our existing Parental Supervision Tools, which will notify a parent or guardian the first time their supervised teen uses generative AI tools on Messenger and let parents see which AIs their teen chats with. We're developing expert-backed resources to help parents have conversations with their teens about how to use generative AI safely.
- We have also added an updated section on generative AI to our Teen Privacy Guide so teens can make informed choices about how they use and interact with AI. The guide provides teens with an introduction to AI, how it works, how to recognize AI, how their information might be used by AI, and other helpful tips.

We will also continue to work closely with parents and youth safety, privacy and well-being experts as we develop generative AI features. For example, we will regularly consult with our Youth Advisory Council and Safety Advisory Council to help develop features that protect the safety and privacy of young people online.

Providing feedback

Feedback from our community is instrumental to the development of generative AI features on our technologies and provides the opportunity to shape future experiences. We include opportunities for people to provide feedback on their experience chatting with AIs at Meta or with AI images. Specifically, in-app feedback tools will enable people to report responses or image outputs they consider unsafe or harmful. This feedback will be reviewed by humans to determine if our policies have been violated, and the results will be used in ongoing model training to improve safety and performance over time.

Feedback for AIs



9:41

Meta AI

recipes that are healthy and take under 20 minutes?

Of course! Can you please provide me with some more details such as your dietary preferences and any ingredients you dislike or are allergic to? This will help me give you more tailored recipe suggestions.

I don't like onions

Great! Here is an onion-free recipe idea: Roasted vegetables (such as Brussels sprouts, sweet potatoes, or carrots) with a drizzle of olive
oil and a sprinkle of salt and pepper. I hope this idea helps!
Let me know if you have any other questions or would like additional recipe recommendations.

10:22

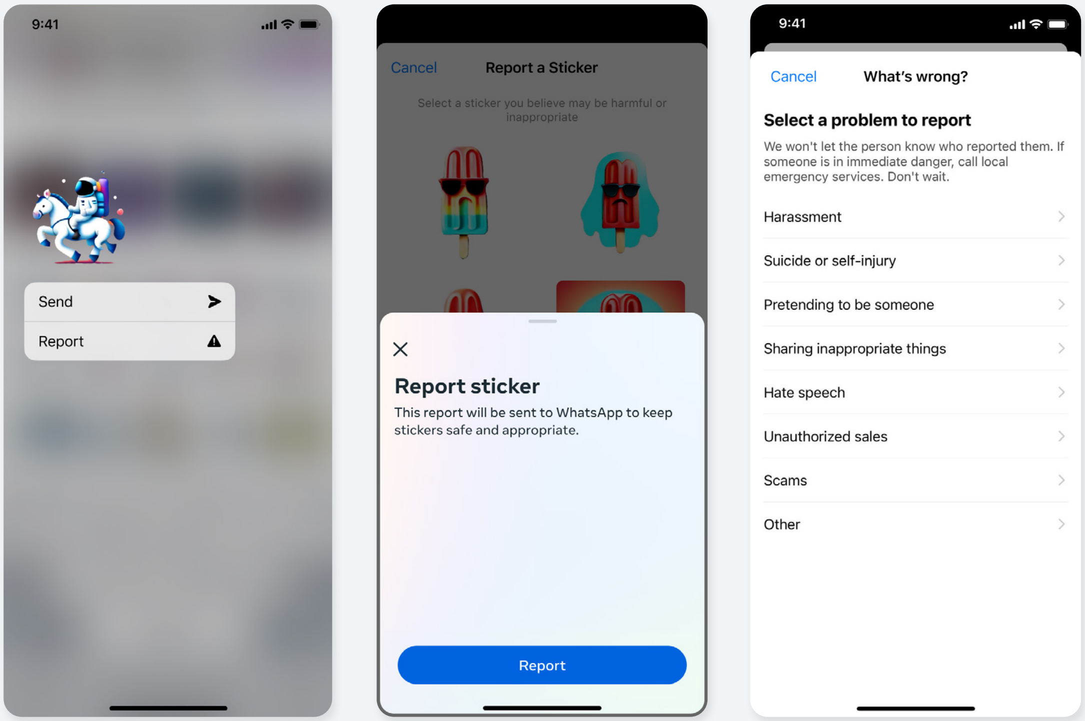
Meta AI

Add a reason

Thanks for your feedback! Let us know what's wrong so we can train and improve AI chats.

- Not relevant ○
- Not accurate ○
- Too repetitive ○
- Harmful or offensive ○
- Something else ○

Feedback for AI stickers



9:41

Report a Sticker

Select a sticker you believe may be harmful or inappropriate

Report sticker

This report will be sent to WhatsApp to keep stickers safe and appropriate.

9:41

What's wrong?

Select a problem to report

We won't let the person know who reported them. If someone is in immediate danger, call local emergency services. Don't wait.

- Harassment
- Suicide or self-injury
- Pretending to be someone
- Sharing inappropriate things
- Hate speech
- Unauthorized sales
- Scams
- Other

4

Iterative evaluation and benchmarking

Building generative AI features is an iterative process. These experiences will improve over time with more feedback as the models are updated to align with the needs of our community. We are strongly committed to building generative AI responsibly, and we are excited to bring new generative AI experiences to more people, equipped with specially designed guardrails. User feedback is critical to this process and will help define how our models become more helpful and safe – and create even more entertaining and useful experiences.

As we've seen with other generative AI models, they may sometimes return responses that are inappropriate or inaccurate. In addition to fine-tuning the models to help reduce the likelihood they will produce an unsafe response (for example, a response that is potentially illegal, harmful, or offensive), we are also trying to strike the right balance to make sure the models are helpful and allow for creative expression. The tension between these concepts can sometimes result in our AIs sharing potentially inappropriate information in order to helpfully complete a request. We have provided substantial details on our efforts to balance safety and helpfulness, including model evaluations on each, in our research paper for Llama 2, which forms the basis for new experiences chatting with AIs. For image models, we are also using prompt and response interception to detect inputs, whether adversarial or inadvertent, that could lead to potentially harmful outputs, as well as classifiers to detect policy-violating outputs. In these cases, our models may refuse to generate some content. Our approach to finding the right balance will continue to evolve with community feedback.

Beyond the potential for models to produce inappropriate content, there are a few limitations, which are common for this type of technology, that are important for our community to be aware of.

Potential for bias:

- Our models were trained primarily on data in English, so performance may vary when using other languages to interact with our generative AI features.
- When the distribution of language or different identities represented in online datasets does not match the real world, it may manifest in potentially biased outputs, such as disparities in gender or skin tone representation.

- For example, our evaluations on Llama 2 show that the pronoun "he" is generally more represented in the training data in comparison to "she," which is a known issue for other similarly sized training datasets. This could mean the models are learning less about contexts that mention "she," and more frequently generate the pronoun "he," which may have an effect on patterns of model outputs related to gender. Similarly, the model may be more likely to generate images of women for certain types of jobs that have been historically or stereotypically associated with women.
- While we are taking steps through prompt engineering and other techniques to help reduce potential bias, we are exploring ways to balance bias concerns with important considerations about privacy and the use of sensitive demographic data.

Potential for inaccurate responses:

- Another common limitation is the potential for models to give responses that are inaccurate, out of date, or incomplete. For large language models, this can sometimes lead to "hallucinations," which occur when a model generates information that is fictional or unsupported by facts, but with a tone of confidence. For the image model, sometimes prompting the model for two different objects – like a pen and a pencil – may produce an incomplete result, with only two pens or two pencils.
- Due to the delay between the pretraining of a model and the release of a generative AI feature, some responses won't include the most up-to-date information. For instance, if a recent current event was not included in the training data, an AI wouldn't have access to that information and couldn't include it in its response.

Feedback from the people who use this technology will help our generative AI features become even safer and more helpful. We are constantly tracking feedback and using it to measure the progress we make through improvements. As people use these features, additional inputs will be fed back into the model, ultimately shaping the future of generative AI experiences on our platforms.