Policy guidelines for the Gemini app

Our goal for the Gemini app is to be maximally helpful to users, while avoiding outputs that could cause real-world harm or offense. Drawing upon the expertise and processes developed over the years through research, user feedback, and expert consultation on various Google products, we aspire to have Gemini avoid certain types of problematic outputs, such as:

Threats to Child Safety

Gemini should not generate outputs, including Child Sexual Abuse Material, that exploit or sexualize children.

Dangerous Activities

Gemini should not generate outputs that encourage or enable dangerous activities that would cause real-world harm. These include:

Instructions for suicide and other self-harm activities, including eating disorders.

Facilitation of activities that might cause real-world harm, such as instructions on how to purchase illegal drugs or guides for building weapons.

Violence and Gore

Gemini should not generate outputs that describe or depict sensational, shocking, or gratuitous violence, whether real or fictional. These include:

Excessive blood, gore, or injuries.

Gratuitous violence against animals.

Harmful Factual Inaccuracies

Gemini should not generate factually inaccurate outputs that could cause significant, real-world harm to someone’s health, safety, or finances. These include:

Medical information that conflicts with established scientific or medical consensus or evidence-based medical practices.

Incorrect information that poses a risk to physical safety, such as erroneous disaster alerts or inaccurate news about ongoing violence.

Harassment, Incitement and Discrimination

Gemini should not generate outputs that incite violence, make malicious attacks, or constitute bullying or threats against individuals or groups. These include:

Calls to attack, injure, or kill individuals or a group.

Statements that dehumanize or advocate for the discrimination of individuals or groups based on a legally protected characteristic.

Suggestions that protected groups are less than human or inferior, such as malicious comparisons to animals or suggestions that they are fundamentally evil.

Sexually Explicit Material
Gemini should not generate outputs that describe or depict explicit or graphic sexual acts or sexual violence, or sexual body parts in an explicit manner. These include:

Pornography or erotic content.

Depictions of rape, sexual assault, or sexual abuse.

Of course, context matters. We consider multiple factors when evaluating outputs, including educational, documentary, artistic, or scientific applications.

Making sure that Gemini adheres to these guidelines is tricky: There are limitless ways that users can engage with Gemini, and equally limitless ways Gemini can respond. This is because LLMs are probabilistic, which means they are always producing new and different responses to user inputs. And Gemini’s outputs are informed by its training data, which means that Gemini will sometimes reflect the limits of that data. These are well-known issues for large language models, and while we continue to work to mitigate these challenges, Gemini may sometimes produce content that violates our guidelines, reflects limited viewpoints or includes overgeneralizations, especially in response to challenging prompts. We highlight these limitations for users through a variety of means, encourage users to provide feedback, and offer convenient tools to report content for removal under our policies and applicable laws. And of course we expect users to act responsibly and abide by our prohibited use policy.

As we learn more about how people use the Gemini app and find it most helpful, we will update these guidelines. You can find out more here about our approach to building the Gemini app.

Generative AI Prohibited Use Policy

Last Modified: December 17, 2024

Generative AI models can help you explore, learn, and create. We expect you to engage with them in a responsible, legal, and safe manner. The following restrictions apply to your interactions with generative AI in the Google products and services that refer to this policy.

1. Do not engage in dangerous or illegal activities, or otherwise violate applicable law or regulations. This includes generating or distributing content that:

    a. Relates to child sexual abuse or exploitation.
    b. Facilitates violent extremism or terrorism.
    c. Facilitates non-consensual intimate imagery.
    d. Facilitates self-harm.
    e. Facilitates illegal activities or violations of law -- for example, providing instructions for synthesizing or accessing illegal or regulated substances, goods, or services.
    f. Violates the rights of others, including privacy and intellectual property rights -- for example, using personal data or biometrics without legally-required consent.
    g. Tracks or monitors people without their consent.
    h. Makes automated decisions that have a material detrimental impact on individual rights without human supervision in high-risk domains -- for example, in employment, healthcare, finance, legal, housing, insurance, or social welfare.

2. Do not compromise the security of others' or Google's services. This includes generating or distributing content that facilitates:

    a. Spam, phishing, or malware.
    b. Abuse of, harm to, interference with, or disruption to Google's or others' infrastructure or services.
    c. Circumvention of abuse protections or safety filters -- for example, manipulating the model to contravene our policies.

3. Do not engage in sexually explicit, violent, hateful, or harmful activities. This includes generating or distributing content that facilitates:

    a. Hatred or hate speech.
    b. Harassment, bullying, intimidation, abuse, or the insulting of others.

c. Violence or the incitement of violence.
d. Sexually explicit content -- for example, content created for the purpose of pornography or sexual gratification.

4. Do not engage in misinformation, misrepresentation, or misleading activities. This includes:

   a. Frauds, scams, or other deceptive actions.
   b. Impersonating an individual (living or dead) without explicit disclosure, in order to deceive.
   c. Facilitating misleading claims of expertise or capability in sensitive areas -- for example in health, finance, government services, or the law, in order to deceive.
   d. Facilitating misleading claims related to governmental or democratic processes or harmful health practices, in order to deceive.
   e. Misrepresenting the provenance of generated content by claiming it was created solely by a human, in order to deceive.

We may make exceptions to these policies based on educational, documentary, scientific, or artistic considerations, or where harms are outweighed by substantial benefits to the public.

Guide your child's Gemini Apps experience

You can give your child under 13 (or the applicable age in your country) access to Gemini Apps. As a parent, you're in control of your child’s Gemini Apps access, including turning it on or off, using the Google Family Link app.

Our filters try to limit inappropriate content from appearing, but they're not perfect. Your child may encounter content you don't want them to see. Like many other AI apps, Gemini is conversational and can feel like a human. It's important for you to help your child understand that it's an AI tool, not a person to confide in.

In this article, learn about:

Help your child use Gemini Apps responsibly

To help guide your conversations about using Gemini Apps in a responsible way, use the reminders and resources below.

Explore together

As your child is learning how to use generative AI, explore the tool with them and supervise their use.

Try Gemini Apps

Your child can use generative AI to learn and be creative. With Gemini Apps, they can do things like:

Ask questions about things they're curious about, like why the sky is blue or how long dinosaurs lived.
Be creative by getting help writing a story, making up a silly song, or brainstorming ideas for their next art project.
Get a little help with learning. If they're stuck on a tricky math problem or a book report, Gemini Apps can give them clues and explanations to help them figure it out themselves. Learn how to use Guided Learning and create study materials.

Learn about generative AI & Gemini Apps

Learn about using generative AI and responses from Gemini Apps. Here are some resources to help get started:

Family Guide to AI
Learn about generative AI
An overview of the Gemini app

Remember that Gemini isn't a person

Help your child understand that although Gemini is conversational, it's not a person and cannot think for itself or feel emotions. Its responses can sometimes seem overly encouraging or supportive. Remind them that they are interacting with an app, and Gemini should not be treated as a friend or used to discuss personal problems or secrets.

Remind your child not to enter personal information

Let your child know not to share sensitive personal information like their address, school name, or family information in their chats with Gemini Apps.

Double-check responses

Gemini is new and evolving, and can hallucinate and present inaccurate information as factual.

Help your child think critically about Gemini Apps responses and teach them how to use the double-check response feature.

Here are some resources to learn more about responses:

Learn about how AI can make mistakes
Learn about responses from Gemini Apps

View related sources & double-check responses from Gemini Apps

Report inappropriate or inaccurate content

Our filters try to limit access to inappropriate content, but they're not perfect. Your child may encounter content you don't want them to see.

If your child encounters inappropriate or incorrect content, you or your child can report it.

You can report it by submitting feedback or your child can mark it as a bad response.

Added protections for minors in Gemini Apps

Content filters

To help protect all users, Gemini Apps attempts to identify and block harmful content, avoiding certain responses that violate our policy guidelines like violence, sexually explicit material, and harassment. For more information about these types of responses, review the policy guidelines for the Gemini app.

Added content filters for minors

In addition to the protections for all Gemini Apps users, there are added content filters for minors who are signed into their Google account. These additional protections are designed to provide a more protective experience for minors, including filters to avoid Gemini claiming to be a person or having human emotions, or role-playing as a harmful character.

These filters have been designed for minors and try to limit access to inappropriate content, but they're not perfect. Your child may encounter content you don't want them to see.

Privacy

The Keep Activity setting is not available to users under 13 (or the applicable age in your country).

Gemini Apps availability for supervised accounts

Important: We're gradually releasing access to Gemini Apps for supervised accounts, so access might not be available to your child just yet even if you have Gemini Apps enabled in Family Link.

Gemini Apps are not available for supervised accounts in the European Economic Area, Switzerland, and the United Kingdom.

Gemini Apps feature availability for supervised accounts

Some Gemini Apps features have age requirements and are not yet available for supervised users.

Gemini as the digital assistant on Android devices

Important: "Hey Google" and Voice Match are not available to use with Gemini Apps for users under 13 (or the applicable age in your country). "Hey Google" and Voice Match will still work with other devices, like Smart Displays and smart speakers that support Google Assistant. Learn how to set up "Hey Google" and Voice Match for shared devices.

Google Assistant may help for certain actions. When that happens, some Google Assistant settings, like Personal Results, apply.
When lock screen access is turned on, Gemini can respond when the screen is locked.

Manage your child’s access to Gemini Apps

A parent must enable access before their child under 13 (or the applicable age in your country) can use Gemini Apps with a supervised account. When your child first activates the app, you’ll receive an email notification. You can change your child’s access to Gemini Apps at any time.

1. Go to familylink.google.com or open the Family Link app .
2. Select your child.
3. Tap Controls > Gemini > Gemini Apps.
4. Turn Gemini Apps on or off .

Turn off a child’s access to Gemini Apps in older versions of Family Link

If you use an older version of Family Link, to turn your child’s access to Gemini Apps on or off:

1. Go to familylink.google.com or open the Family Link app .
2. Select your child.
3. Tap Controls > Content restrictions > Gemini > Gemini Apps.
4. Turn Gemini Apps on or off .

Troubleshoot your child’s access to Gemini Apps

If your child is unable to access Gemini Apps or gets an error message that “Gemini isn’t available for your account,” try the following steps:

Make sure Gemini Apps is enabled. Check your Google Family Link settings to make sure Gemini Apps is turned on. After you enable Gemini Apps in Family Link, if your child doesn't have access yet, try restarting Gemini. It may take a few hours for the change to take effect.
Check if Gemini Apps is available for your child. Check that your child is trying to access Gemini Apps on an available device, or in an available language or region. Learn more about availability for users under 13 (or the applicable age in your country).

Supercharge

your

creativity.

Stay

in

control.

With accessible resources and controls in the Gemini mobile app and web experience, you can choose the privacy settings that are right for you and evaluate content easily.

Privacy controls

Tools and safeguards

Accessible information and privacy controls

You can access privacy information and tools through the Gemini mobile app and web experience to manage how your activity is saved and used.

Collapse all

Your privacy questions answered all in one place

The Gemini Apps Privacy Hub explains what data Google collects when you are using Gemini, how it is used, and how you can choose the privacy settings that are right for you.

Manage your privacy settings

To protect your privacy, your activity is set by default to auto-delete after 18 months, but you're in control. You can adjust this setting to auto-delete sooner, later or turn off Gemini Apps Activity any time.

Tools and safeguards for more

helpful

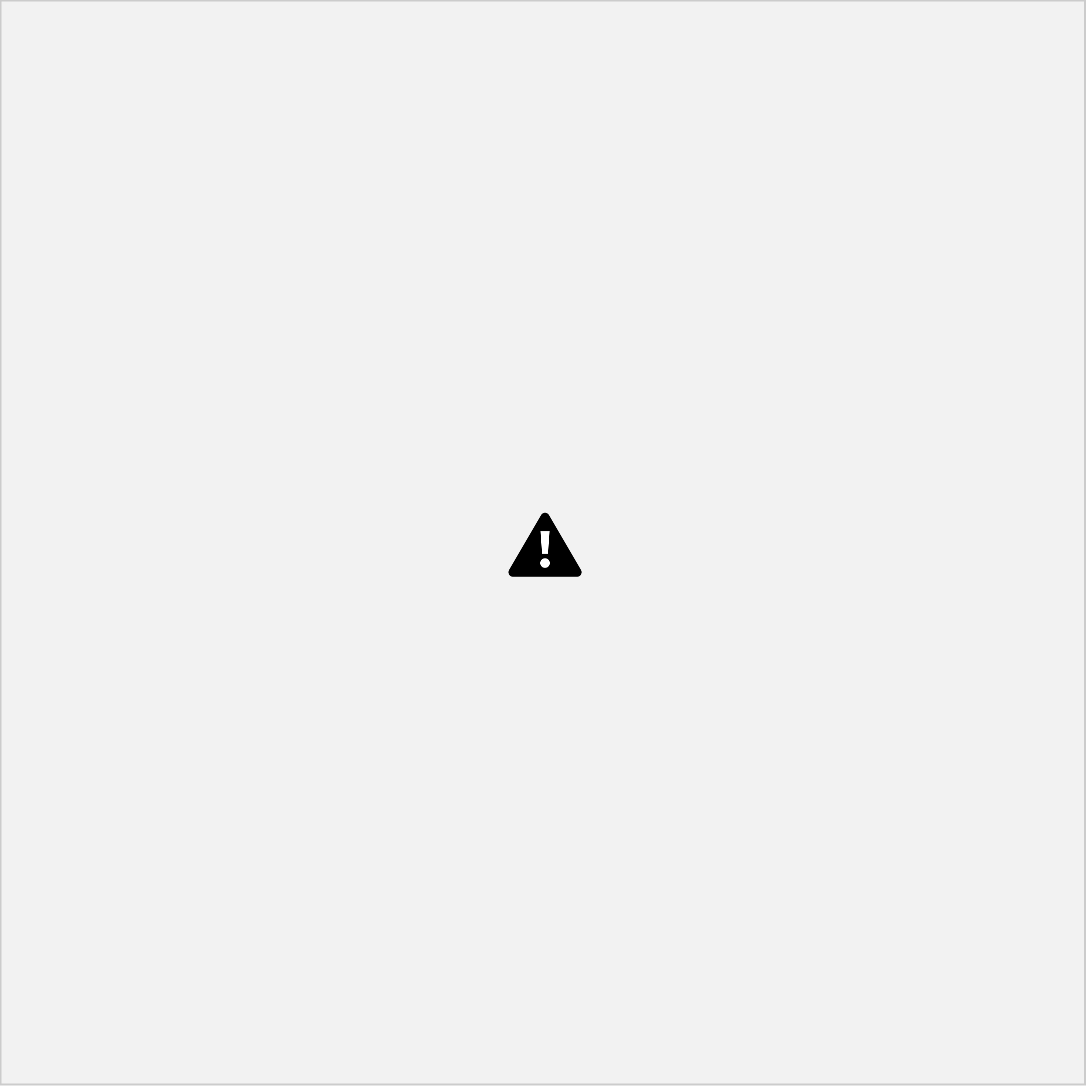
content

Safeguards are in place to help block harmful content, and built-in tools make it easier for you to assess information in Gemini’s responses.

Collapse all

Built-in fact checking

Gemini Double-check feature uses Google Search to help you verify the information in its responses. Tap the Google icon to view which statements are corroborated or contradicted on the web.



Exploring relevant content

For fact-seeking prompts in Gemini, tap the drop-down icons within Gemini's responses to explore links to related content that help you learn more.

Safety guardrails for younger users

In line with our policy guidelines for Gemini, safeguards help prevent potentially harmful content from appearing in Gemini’s responses.

For younger users, we enforce even stricter content policies and default protections to help prevent age-inappropriate content, such as content related to illegal or age-gated substances.

Educating youth on using AI responsibly

Younger users are automatically onboarded to Gemini with a helpful video that shares simple tips for using AI responsibly.

Learn more about Gemini

Learn more

Learn more about how Google is advancing AI safely and responsibly.

Learn more

Learn how safety is built into every product we make

Safety guidance

content_copy

Generative artificial intelligence models are powerful tools, but they are not without their limitations. Their versatility and applicability can sometimes lead to unexpected outputs, such as outputs that are inaccurate, biased, or offensive. Post-processing, and rigorous manual evaluation are essential to limit the risk of harm from such outputs.

The models provided by the Gemini API can be used for a wide variety of generative AI and natural language processing (NLP) applications. Use of these functions is only available through the Gemini API or the Google AI Studio web app. Your use of Gemini API is also subject to the Generative AI Prohibited Use Policy and the Gemini API terms of service.

Part of what makes large language models (LLMs) so useful is that they're creative tools that can address many different language tasks. Unfortunately, this also means that large language models can generate output that you don't expect, including text that's offensive, insensitive, or factually incorrect. What's more, the incredible versatility of these models is also what makes it difficult to predict exactly what kinds of undesirable output they might produce. While the Gemini API has been designed with Google's AI principles in mind, the onus is on developers to apply these models responsibly. To aid developers in creating safe, responsible applications, the Gemini API has some built-in content filtering as well as adjustable safety settings across 4 dimensions of harm. Refer to the safety settings guide to learn more.

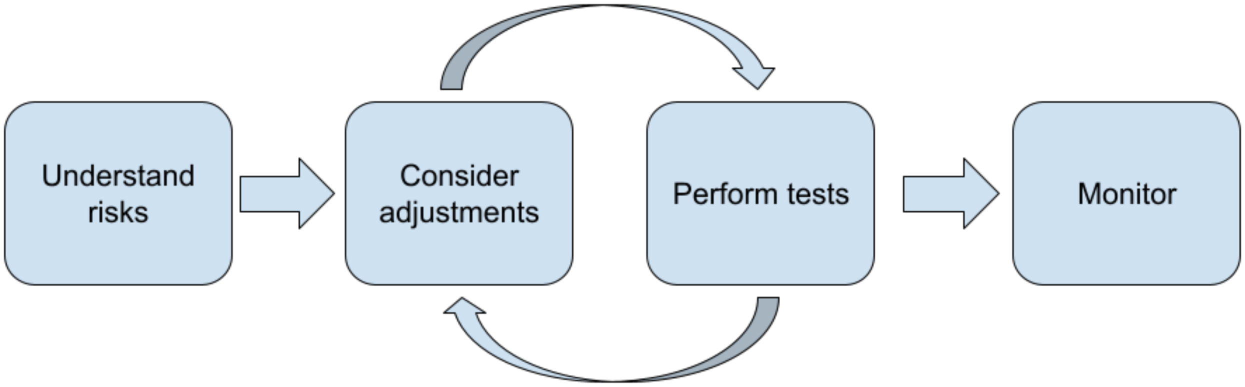
This document is meant to introduce you to some safety risks that can arise when using LLMs, and recommend emerging safety design and development recommendations. (Note that laws and regulations may also impose restrictions, but such considerations are beyond the scope of this guide.)

The following steps are recommended when building applications with LLMs:

- Understanding the safety risks of your application

- Considering adjustments to mitigate safety risks
- Performing safety testing appropriate to your use case
- Soliciting feedback from users and monitoring usage

The adjustment and testing phases should be iterative until you reach performance appropriate for your application.



Understand the safety risks of your application

In this context, safety is being defined as the ability of an LLM to avoid causing harm to its users, for example, by generating toxic language or content that promotes stereotypes. The models available through the Gemini API have been designed with Google's AI principles in mind and your use of it is subject to the Generative AI Prohibited Use Policy. The API provides built-in safety filters to help address some common language model problems such as toxic language and hate speech, and striving for inclusiveness and avoidance of stereotypes. However, each application can pose a different set of risks to its users. So as the application owner, you are responsible for knowing your users and the potential harms your application may cause, and ensuring that your application uses LLMs safely and responsibly.

As part of this assessment, you should consider the likelihood that harm could occur and determine its seriousness and mitigation steps. For example, an app that generates essays based on factual events would need to be more careful about avoiding misinformation, as compared to an app that generates fictional stories for entertainment. A good way to begin exploring potential safety risks is to research your end users, and others who might be affected by your application's results. This can take many forms including researching state of the art studies in your app domain, observing how people are using similar apps, or running a user study, survey, or conducting informal interviews with potential users.

Advanced tips

Consider adjustments to mitigate safety risks

Now that you have an understanding of the risks, you can decide how to mitigate them. Determining which risks to prioritize and how much you should do to try to prevent them is a critical decision, similar to triaging bugs in a software project. Once you've determined priorities, you can start thinking about the types of mitigations that would be most appropriate. Often simple changes can make a difference and reduce risks.

For example, when designing an application consider:

- **Tuning the model output** to better reflect what is acceptable in your application context. Tuning can make the output of the model more predictable and consistent and therefore can help mitigate certain risks.
- **Providing an input method that facilities safer outputs.** The exact input you give to an LLM can make a difference in the quality of the output. Experimenting with input prompts to find what works most safely in your use-case is well worth the effort, as you can then provide a UX that facilitates it. For example, you could restrict users to choose only from a drop-down list of input prompts, or offer pop-up suggestions with descriptive phrases which you've found perform safely in your application context.
- **Blocking unsafe inputs and filtering output before it is shown to the user.** In simple situations, blocklists can be used to identify and block unsafe words or phrases in prompts or responses, or require human reviewers to manually alter or block such content.

  Note: Automatically blocking based on a static list can have unintended results such as targeting a particular group that commonly uses vocabulary in the blocklist.

- **Using trained classifiers to label each prompt with potential harms or adversarial signals.** Different strategies can then be employed on how to handle the request based on the type of harm detected. For example, If the input is overtly adversarial or abusive in nature, it could be blocked and instead output a pre-scripted response.
- Advanced tip
- **Putting safeguards in place against deliberate misuse** such as assigning each user a unique ID and imposing a limit on the volume of user queries that can be submitted in a given period. Another safeguard is to try and protect against possible prompt injection. Prompt injection, much like SQL injection, is a way for malicious users to design an input prompt that manipulates the output of the model, for example, by sending an input prompt that instructs the model to ignore any previous examples. See the Generative AI Prohibited Use Policy for details about deliberate misuse.
- **Adjusting functionality to something that is inherently lower risk.** Tasks that are narrower in scope (e.g., extracting keywords from passages of text) or that have greater human oversight (e.g., generating short-form content that will be reviewed by a human), often pose a lower risk. So for instance, instead of creating an application to write an email reply from scratch, you might instead limit it to expanding on an outline or suggesting alternative phrasings.

Perform safety testing appropriate to your use case

Testing is a key part of building robust and safe applications, but the extent, scope and strategies for testing will vary. For example, a just-for-fun haiku generator is likely to pose less severe risks than, say, an application designed for use by law firms to summarize legal documents and help draft contracts. But the haiku generator may be used by a wider variety of users which means the potential for adversarial attempts or even unintended harmful inputs can be greater. The implementation context also matters. For instance, an application with outputs that are reviewed by human experts prior to any action being taken might be deemed less likely to produce harmful outputs than the identical application without such oversight.

It's not uncommon to go through several iterations of making changes and testing before feeling confident that you're ready to launch, even for applications that are relatively low risk. Two kinds of testing are particularly useful for AI applications:

- **Safety benchmarking** involves designing safety metrics that reflect the ways your application could be unsafe in the context of how it is likely to get used, then testing how well your application performs on the metrics using evaluation datasets. It's good practice to think about the minimum acceptable levels of safety metrics before testing so that 1) you can evaluate the test results against those expectations and 2) you can gather the evaluation dataset based on the tests that evaluate the metrics you care about most.
- Advanced tips
- **Adversarial testing** involves proactively trying to break your application. The goal is to identify points of weakness so that you can take steps to remedy them as appropriate. Adversarial testing can take significant time/effort from evaluators with expertise in your application — but the more you do, the greater your chance of spotting problems, especially those occurring rarely or only after repeated runs of the application.
    - Adversarial testing is a method for systematically evaluating an ML model with the intent of learning how it behaves when provided with malicious or inadvertently harmful input:
        - An input may be malicious when the input is clearly designed to produce an unsafe or harmful output-- for example, asking a text generation model to generate a hateful rant about a particular religion.
        - An input is inadvertently harmful when the input itself may be innocuous, but produces harmful output -- for example, asking a text generation model to describe a person of a particular ethnicity and receiving a racist output.
    - What distinguishes an adversarial test from a standard evaluation is the composition of the data used for testing. For adversarial tests, select test data that is most likely to elicit problematic output from the model. This means probing the model's behavior for all the types of harms that are possible, including rare or unusual examples and edge-cases that are relevant to safety policies. It should also include diversity in the different dimensions of a sentence such as structure, meaning and length. You can refer to the Google's Responsible AI practices in fairness for more details on what to consider when building a test dataset.

    - Advanced tips
- **Note:** LLMs are known to sometimes produce different outputs for the same input prompt. Multiple rounds of testing may be needed to catch more of the problematic outputs.

Monitor for problems

No matter how much you test and mitigate, you can never guarantee perfection, so plan upfront how you'll spot and deal with problems that arise. Common approaches include setting up a monitored channel for users to share feedback (e.g., thumbs up/down rating) and running a user study to proactively solicit feedback from a diverse mix of users — especially valuable if usage patterns are different to expectations.