An overview of the Gemini app

James Manyika, SVP, Research, Technology and Society, and **Sissie Hsiao**, Vice President and General Manager, Google Assistant and Gemini App

Editor's note: This is a living document and will be updated periodically as we continue to rapidly improve the Gemini app's capabilities as well as address the limitations inherent to LLMs. This overview was last updated on July 25, 2024. For the latest updates on the Gemini app, visit the Release Updates log or read more on the Google Keyword blog.

We have long seen the potential of AI to make information and computing more accessible and useful to people. We have made pioneering advances on large language models (LLMs) and have seen great progress across Google and in this field more broadly. For several years, we have applied LLMs in the background to improve many of our products, such as autocompleting sentences in Gmail, expanding Google Translate, and helping us better understand queries in Google Search. We continue using LLMs for many Google services, as well as to power the Gemini app, which allows people to collaborate directly with generative AI. We want the Gemini app to be the most helpful and personal AI assistant, giving users direct access to Google's latest AI models.

While we're at an important inflection point and encouraged by the widespread excitement around generative AI, it's still early days for this technology. This explainer outlines how we're approaching our work on the Gemini app ("Gemini"), including its mobile and web experiences — what it is, how it works and its current capabilities and limitations. Our approach to building Gemini will evolve as its underlying technology evolves, and as we learn from ongoing research, experience and user feedback.

What Gemini is

Gemini is an interface to a multimodal LLM (handling text, audio, images and more). Gemini is based on Google's cutting-edge research in LLMs, which began with the Word2Vec paper in 2013 that proposed novel model architectures that mapped words as mathematical concepts, followed by the introduction of a neural conversational model in 2015. This framework demonstrated how models could predict the next sentence in a conversation based on the previous sentence or sentences, leading to more natural conversational experiences. This was followed by our breakthrough work on Transformer in 2017 and multi-turn chat capabilities in 2020, which demonstrated even more compelling generative language progress.

We initially launched Gemini (then called Bard) as an experiment in March 2023 in accordance with our AI Principles. Since then, users have turned to Gemini to write compelling emails, debug tricky coding problems, brainstorm ideas for upcoming events, get help learning difficult concepts, and so much more. Today, Gemini is a versatile AI tool that can help you in many ways. We already see Gemini helping people be more productive, more creative, and more curious and we add new functionality and innovations regularly.

Productivity

For starters, Gemini can save you time. For example, say you are looking to summarize a long research document; Gemini lets you upload it and gives you a useful synthesis. Gemini can also help with coding tasks, and coding has quickly become one of its most popular applications.

Creativity

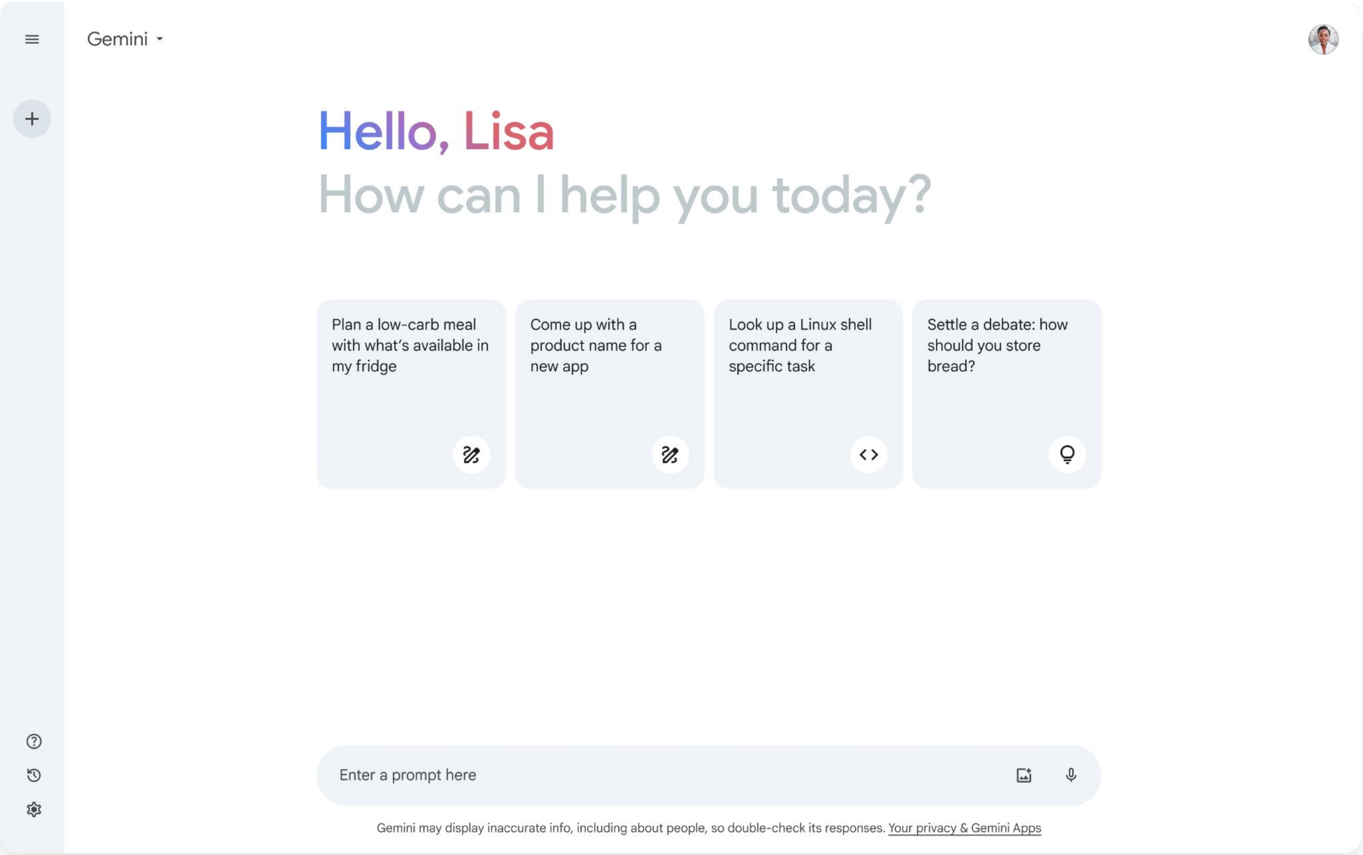
Gemini can also help bring your ideas to life and spark your creativity. For example, if you're writing a blog post, Gemini can create an outline and generate images that help illustrate your post. And coming soon with Gems, you will be able to customize Gemini with specific instructions and have it act as a subject matter expert to help you accomplish your personal goals.

Curiosity

Gemini can be a jumping off point for exploring your ideas and things you'd like to learn more about. For instance, it can explain a complex concept simply or surface relevant insights on a topic or image. And soon, it will pair these insights with recommended content from across the web to learn more about specific topics.

Gemini's capabilities are rapidly expanding -- soon, you'll be able to point your phone's camera at an object, say, for example, the Golden Gate bridge and ask Gemini to tell you about its paint color (if you're wondering, it's "International Orange"). You'll also be able to ask Gemini to help you navigate a restaurant's menu in another language and recommend a dish you're likely to enjoy. These are just two examples of the new capabilities coming soon to Gemini.

Of course we rigorously train and monitor Gemini so that its responses are likely to be reliable and aligned with your expectations. We also talk with industry experts, educators, policymakers, business leaders, civil and human rights leaders, and content creators to explore new applications, risks, and limitations of this emerging technology.



Gemini

Hello, Lisa
How can I help you today?

Plan a low-carb meal with what's available in my fridge

Come up with a product name for a new app

Look up a Linux shell command for a specific task

Settle a debate: how should you store bread?

Enter a prompt here

Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy & Gemini Apps

How Gemini works

Pre-training

Gemini is powered by Google's most capable AI models, designed with varying capabilities and use cases. Like most LLMs today, these models are pre-trained on a variety of data from publicly available sources. We apply quality filters to all datasets, using both heuristic rules and model-based classifiers. We also perform safety filtering to remove content likely to produce policy-violating outputs. To maintain the integrity of model evaluations, we search for and remove any evaluation data that may have been in our training corpus before using data for training. The final data mixtures and weights are determined through ablations on smaller models. We stage training to alter the mixture composition during training – increasing the weight of domain-relevant data towards the end of training. Data quality can be an important factor for highly performing models, and we believe that many interesting questions remain around finding the optimal dataset distribution for pre-training.

This pre-training allows the model to learn to pick up on patterns in language and use them to predict the next probable word or words in a sequence. For example, as an LLM learns, it can predict that the next word in "peanut butter and ___" is more likely to be "jelly" than "shoelace." However, if an LLM picks only the most probable next word, it will lead to less creative responses. So LLMs are often given flexibility to pick from reasonable, albeit slightly less probable, choices (say, "banana") in order to generate more interesting responses. It's worth noting that while LLMs can perform well on factual prompts and create the impression of retrieving information, they are neither information databases nor deterministic information retrieval systems. So while you can expect a consistent response to a database query (one that is a literal retrieval of the fixed information stored in the database), an LLM's response to the same prompt will not necessarily be the same every time (nor will it literally retrieve the information it was trained on). This is also an important reason why LLMs can generate plausible-sounding responses that can at times contain factual errors — not ideal when factuality matters but potentially useful for generating creative or unexpected outputs.

Post-training

After the initial training, LLMs go through additional steps to refine their responses. One of these is called Supervised Fine-Tuning (SFT), which trains the model on carefully selected examples of excellent answers. It's like teaching children to write by showing them well-written stories and essays.

Next comes Reinforcement Learning from Human Feedback (RLHF). Here, the model learns to generate even better responses based on scores or feedback from a special Reward Model. This Reward Model is trained on human preference data, where responses have been rated relative to one another, teaching it what people prefer. Preference data may sometimes include and expose models to offensive or incorrect data so that they learn how to recognize it and avoid it. You can think of preference data like rewarding a child for a job well done; the model is rewarded for creating answers that people like.

Throughout these stages, it's important to use high-quality training data. Examples used for SFT are typically either written by experts or generated by a model and reviewed by experts.

While these techniques are powerful, they have limitations. For example, even with the Reward Model's help, a given response might not always be perfect. Still, the LLM is optimized to produce the most widely preferred responses based on the feedback it receives, similar to students learning from their teachers' comments.

Responses to user prompts

Response generation is similar to how a human might brainstorm different approaches to answering a question. Once a user provides a prompt, Gemini uses the post-trained LLM, the context in the prompt and the interaction with the user to draft several versions of a response. It also relies on external sources such as Google Search, and/or one of its several extensions, and recently uploaded files (Gemini Advanced only) to generate its responses. This process is known as retrieval augmentation. Given a prompt, Gemini strives to retrieve the most pertinent information from these external sources (e.g., Google Search) and represent them accurately in its response. Augmenting LLMs with external tools is an active area of research. There are a number of ways errors can be introduced, including the query Gemini uses to invoke these external tools, how Gemini interprets the results returned by the tools, and the manner in which these returned results are used to generate the final response. Due to this, responses generated by Gemini should not reflect on the performance of the individual tools used to create that response.

Lastly, before the final response is displayed, each potential response undergoes a safety check to ensure it adheres to predetermined policy guidelines. This process provides a double-check to filter out harmful or offensive information. The remaining responses are then ranked based on their quality, with the highest-scoring version(s) presented back to the user.

We also watermark Gemini text and image outputs using SynthID, our industry-leading digital toolkit for watermarking AI-generated content. For generated images, SynthID adds a digital watermark (one that's imperceptible to the human eye) directly into the pixels. SynthID is an important building block for developing more reliable AI identification tools and can help people make informed decisions about how they interact with AI-generated content.

Human feedback and evaluation

Even with safety checks, some errors may occur. And Gemini responses may not always fully meet your expectations. That's where human feedback comes in. Evaluators assess the quality of responses, identifying areas for improvement and suggesting solutions. This feedback becomes part of the Gemini learning process, described in the "Post-training" section above.

Known limitations of LLM-based interfaces like Gemini

Gemini is just one part of our continuing effort to develop LLMs responsibly. Throughout the course of this work, we have discovered and discussed several limitations associated with LLMs. Here, we focus on six areas of continuing research: **Accuracy:** Gemini's responses might be inaccurate, especially when it's asked about complex or factual topics; **Bias:** Gemini's responses might reflect biases present in its training data; **Multiple Perspectives:** Gemini's responses might fail to show a range of views; **Persona:** Gemini's responses might incorrectly suggest it has personal opinions or feelings, **False positives and false negatives:** Gemini might not respond to some appropriate prompts and provide inappropriate responses to others, and **Vulnerability** to adversarial prompting: users will find ways to stress test Gemini with nonsensical prompts or questions rarely asked in the real world. We continue to explore new approaches and areas for improved performance in each of these areas.

Accuracy

Gemini is grounded in Google's understanding of authoritative information, and is trained to generate responses that are relevant to the context of your prompt and in line with what you're looking for. But like all LLMs, Gemini can sometimes confidently and convincingly generate responses that contain inaccurate or misleading information.

Since LLMs work by predicting the next word or sequences of words, they are not yet fully capable of distinguishing between accurate and inaccurate information on their own. We have seen Gemini present responses that contain or even invent inaccurate information (e.g., misrepresenting how it was trained or suggesting the name of a book that doesn't exist). In response we have created features like "double check", which uses Google Search to find content that helps you assess Gemini's responses, and gives you links to sources to help you corroborate the information you get from Gemini.

Bias

Training data, including from publicly available sources, reflects a diversity of perspectives and opinions. We continue to research how to use this data in a way that ensures that an LLM's response incorporates a wide range of viewpoints, while minimizing inaccurate overgeneralizations and biases.

Gaps, biases, and overgeneralizations in training data can be reflected in a model's outputs as it tries to predict likely responses to a prompt. We see these issues manifest in a number of ways (e.g., responses that reflect only one culture or demographic, reference problematic overgeneralizations, exhibit gender, religious, or ethnic biases, or promote only one point of view). For some topics, there are data voids — in other words, not enough reliable information about a given subject for the LLM to learn about it and then make good predictions — which can result in low-quality or inaccurate responses. We continue to work with domain experts and a diversity of communities to draw on deep expertise outside of Google.

Multiple Perspectives

For subjective topics, Gemini is designed to provide users with multiple perspectives if the user does not request a specific point of view. For example, if prompted for information on something that cannot be verified by primary source facts or authoritative sources — like a subjective opinion on "best" or "worst" — Gemini should respond in a way that reflects a wide range of viewpoints. But since LLMs like Gemini train on the content publicly available on the internet, they can reflect positive or negative views of specific politicians, celebrities, or other public figures, or even incorporate views on just one side of controversial social or political issues. Gemini should not respond in a way that endorses a particular viewpoint on these topics, and we will use feedback on these types of responses to train Gemini to better address them.

Persona

Gemini might at times generate responses that seem to suggest it has opinions or emotions, like love or sadness, since it has trained on language that people use to reflect the human experience. We have developed a set of guidelines around how Gemini might represent itself (i.e., its persona) and continue to finetune the model to provide objective responses.

False positives / negatives

We've put in place a set of policy guidelines to help train Gemini and avoid generating problematic responses. Gemini can sometimes misinterpret these guidelines, producing "false positives" and "false negatives." In a "false positive," Gemini might not provide a response to a reasonable prompt, misinterpreting the prompt as inappropriate; and in a "false negative," Gemini might generate an inappropriate response, despite the guidelines in place. Sometimes, the occurrence of false positives or false negatives may give the impression that Gemini is biased: For example, a false positive might cause Gemini to not respond to a question about one side of an issue, while it will respond to the same question about the other side. We continue to tune these models to better understand and categorize inputs and outputs as language, events and society rapidly evolve.

Vulnerability to adversarial prompting

We expect users to test the limits of what Gemini can do and attempt to break its protections, including trying to get it to divulge its training protocols or other information, or try to get around its safety mechanisms. We have tested and continue to test Gemini rigorously, but we know users will find unique, complex ways to stress-test it further. This is an important part of refining Gemini and we look forward to learning the new prompts users come up with. Indeed, since Gemini launched in 2023, we've seen users challenge it with prompts that range from the philosophical to the nonsensical – and in some cases, we've seen Gemini respond with answers that are equally nonsensical or not aligned with our stated approach. Figuring out methods to help Gemini respond to these sorts of prompts is an on-going challenge and we have continued to expand our internal evaluations and red-teaming to strive toward continued improvement to accuracy, and objectivity and nuance.

How we're continuing to develop Gemini

Application of our Gemini approach

Along with our AI Principles, we recently articulated our approach to our work on Gemini: Gemini should follow your directions, adapt to your needs, and safeguard your experience. Core to our approach is a focus on responsibility and safety. Gemini's policy guidelines seek to avoid certain types of problematic outputs. We are engaging in ongoing adversarial testing with internal "red team" members — product experts and social scientists who intentionally stress test a model to probe it for alignment issues with these policy guidelines and our northstar approach for Gemini — so we can apply what they learn and continuously improve Gemini.

Privacy is also a key consideration as we develop Gemini. The Gemini Apps Privacy Hub has more information about how we build Gemini with privacy by design, and with you in control.

Enabling user and publisher control

We've built a variety of easily accessible Gemini user controls for you to review, update, manage, export, and delete your Gemini data. You can access and review your Gemini prompts, responses, and feedback through the Gemini Apps Activity control. In addition, you can prevent your future Gemini chats from being used to improve Google machine-learning technologies by turning off your Gemini

Apps Activity setting. And like with other Google services, you can also download and export your information through Google's Takeout tool. We also have controls that allow you to manage public links you've created to your Gemini threads, and controls that allow you to turn on/off access to extensions (e.g., Workspace, Maps, YouTube). We're also exploring new ways to give you more control over Gemini's responses, including adjusting filters to enable a broader range of responses.

For publishers, we've launched Google-Extended, a control that web publishers can use to manage whether their sites help improve Gemini and Vertex AI generative APIs. Allowing Google-Extended access to sites' content can help AI models become more accurate and capable over time. In addition to not using the content from opted-out URLs for model training, Gemini will also not use such content for grounding. As AI applications expand, web publishers will face the increasing complexity of managing different uses at scale, and we're committed to engaging with the web and AI communities to explore more machine-readable approaches to choice and control.

Improving Gemini together

We believe in rapid iteration and bringing the best of Gemini to the world. User feedback has accelerated improvements to our models. For example, we use state-of-the-art reinforcement learning techniques to train our models to be more intuitive and imaginative, and to respond with even more quality and accuracy. We continue to invest in research to learn more about the technical, social, and ethical challenges and opportunities of LLMs, both to improve Gemini's model training and tuning techniques as well as to share our learnings with researchers, such as this recent paper on the Ethics of Advanced AI Assistants. We're committed to innovating in this space responsibly, collaborating with users, trusted testers and researchers to find ways for this new technology to benefit the entire ecosystem.

Transparency is important and we are committed to being open about Gemini's development process and limitations. Gemini is not a magical black box; it's constantly evolving and we will continue to share updates on our progress. We've launched a Release Updates page so you can see Gemini's latest features, improvements, and bug fixes, and we will update this overview as appropriate. We will be identifying both where Gemini is useful and helpful, and where we need to continue to iterate and make it better. We are actively adding new capabilities, and through ongoing research, testing, and user feedback, we look forward to improving Gemini together.

Acknowledgments

We appreciate and acknowledge the incredible work of our colleagues on the Gemini app team, Google DeepMind, Trust & Safety, and Google Research.