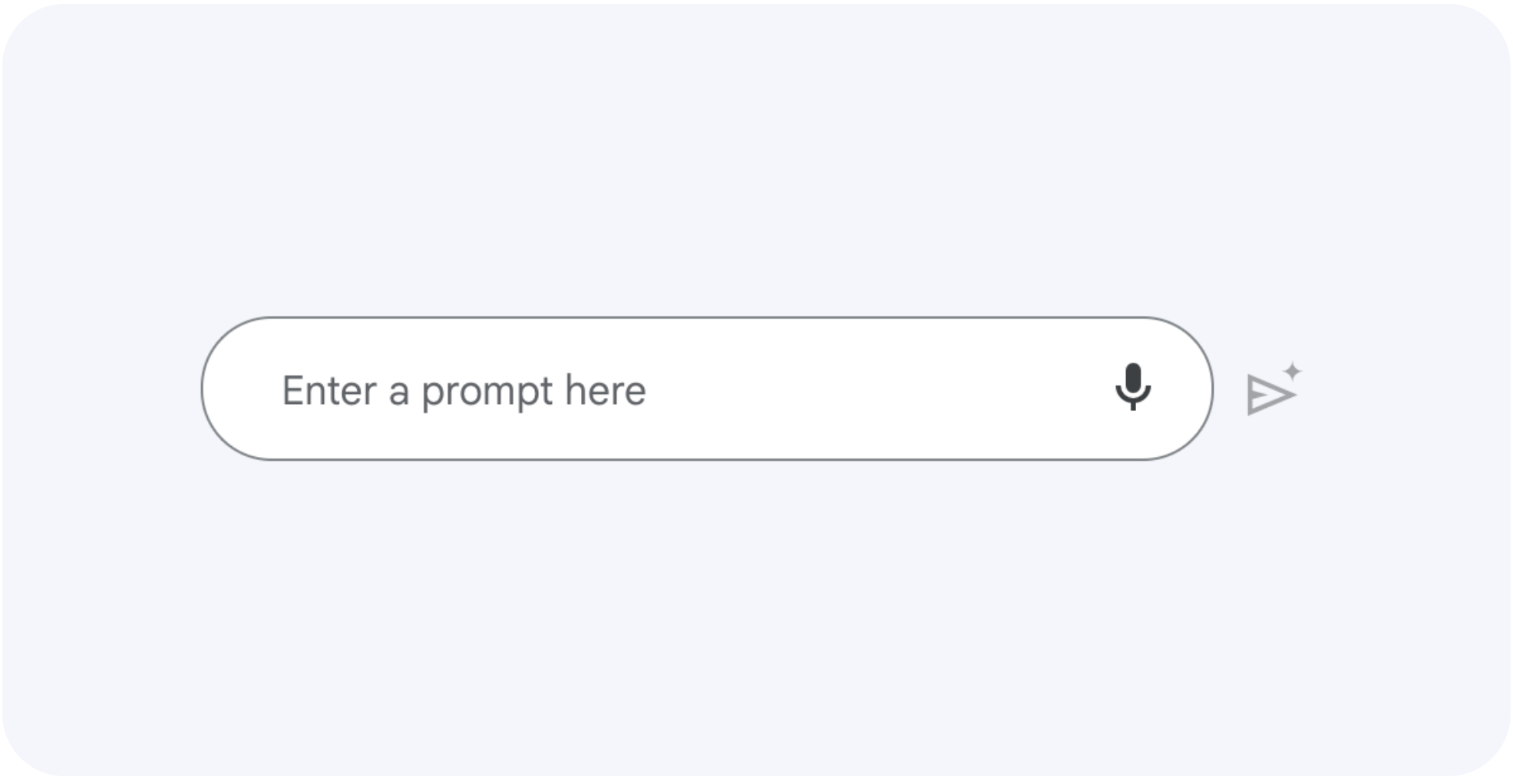
AI Principles Case Study

Responsible Development of

Bard: A Conversational Generative AI Experience



December 2023

8 min.

The thoughtful design behind an AI service supercharging imaginations & boosting productivity.

A helpful creative collaborator

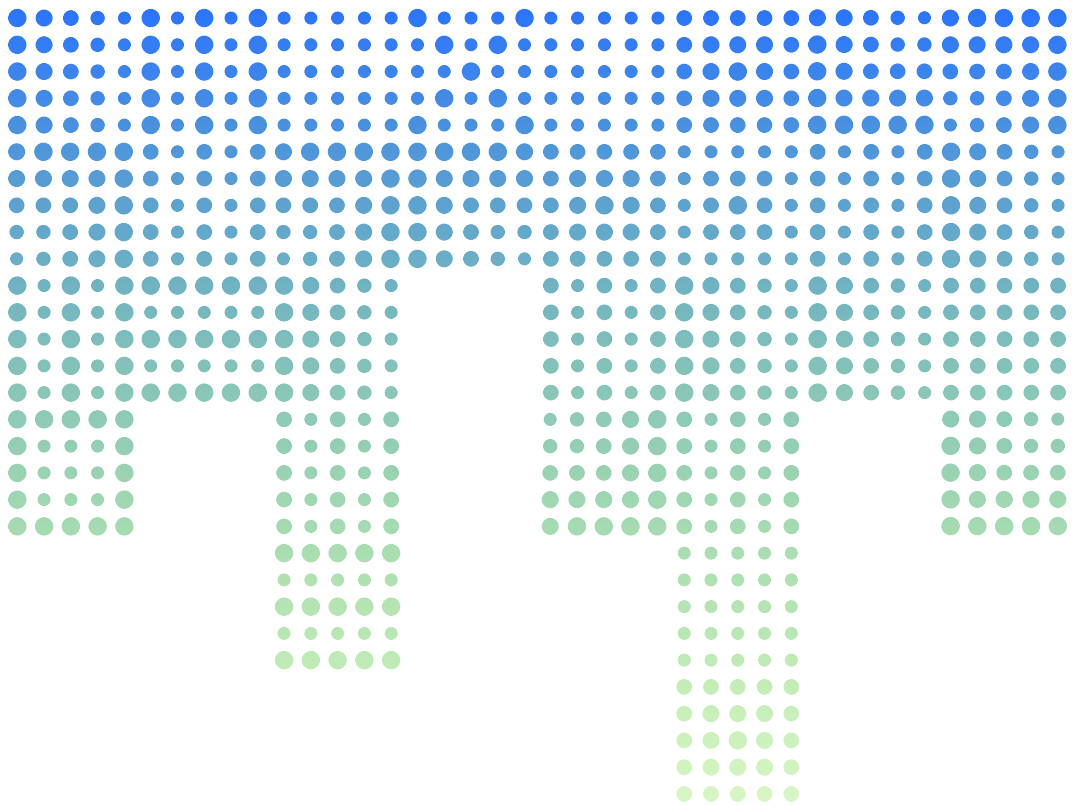
Bard is Google's generative conversational AI experience, launched in early 2023. Generative AI refers to machine learning models that use learned patterns to create entirely new content such as text, images, music and videos, based on human inputs or requests, called "prompts." Bard can support people's productivity, creativity, and curiosity. From planning a party (Bard can come up with a to-do list) to writing a blog post (Bard can provide an outline), people have a new and helpful creative collaborator.

Bard uses a large language model (LLM), Gemini Pro, which has been extensively trained and tested in alignment with Google's AI Principles. Since LLMs are trained on datasets comprising data created and curated by people, the models learn real-world biases. As a result of potential unfair bias in training data, generative AI product output can be harmful, offensive, or factually inaccurate. Much of the challenge in developing Bard was maximizing the product's social benefit while minimizing the potential for harmful or factually inaccurate outputs.

Leading the way

As Google's first standalone publicly available generative AI tool, Bard helped establish many responsible AI precedents at Google, and the Bard team's work on content policies (for example, which types of content Bard is and is not allowed to generate) influenced our current company-wide content policy for generative AI models. The team's thoughtful approach to development also shaped our understanding of emerging best practices for responsible generative AI development, including conducting adversarial testing and providing people with simple, helpful explanations.

Bard was launched as an experiment at first, so that people could access the technology while the Bard team could learn from real-world use by trusted testers from diverse backgrounds and make adjustments as needed. Bard underwent extensive adversarial testing before launch to identify harmful outputs and make model improvements. Bard continues to undergo regular adversarial testing, especially as new features are added. The Bard interface lets people know that they're interacting with AI and reminds them to check its responses. Further, people can give feedback on Bard's responses using the "thumbs up" and "thumbs down" feature.



Bringing ideas to life, responsibly

With new features launching all the time, it remains critical that Bard prioritize factuality, explainability, fairness, and safety.

Google AI Principles guiding the Bard team:

- **AI Principle # 2 (Avoid creating or reinforcing unfair bias) & AI Principle # 3 (Be built and tested for safety)**

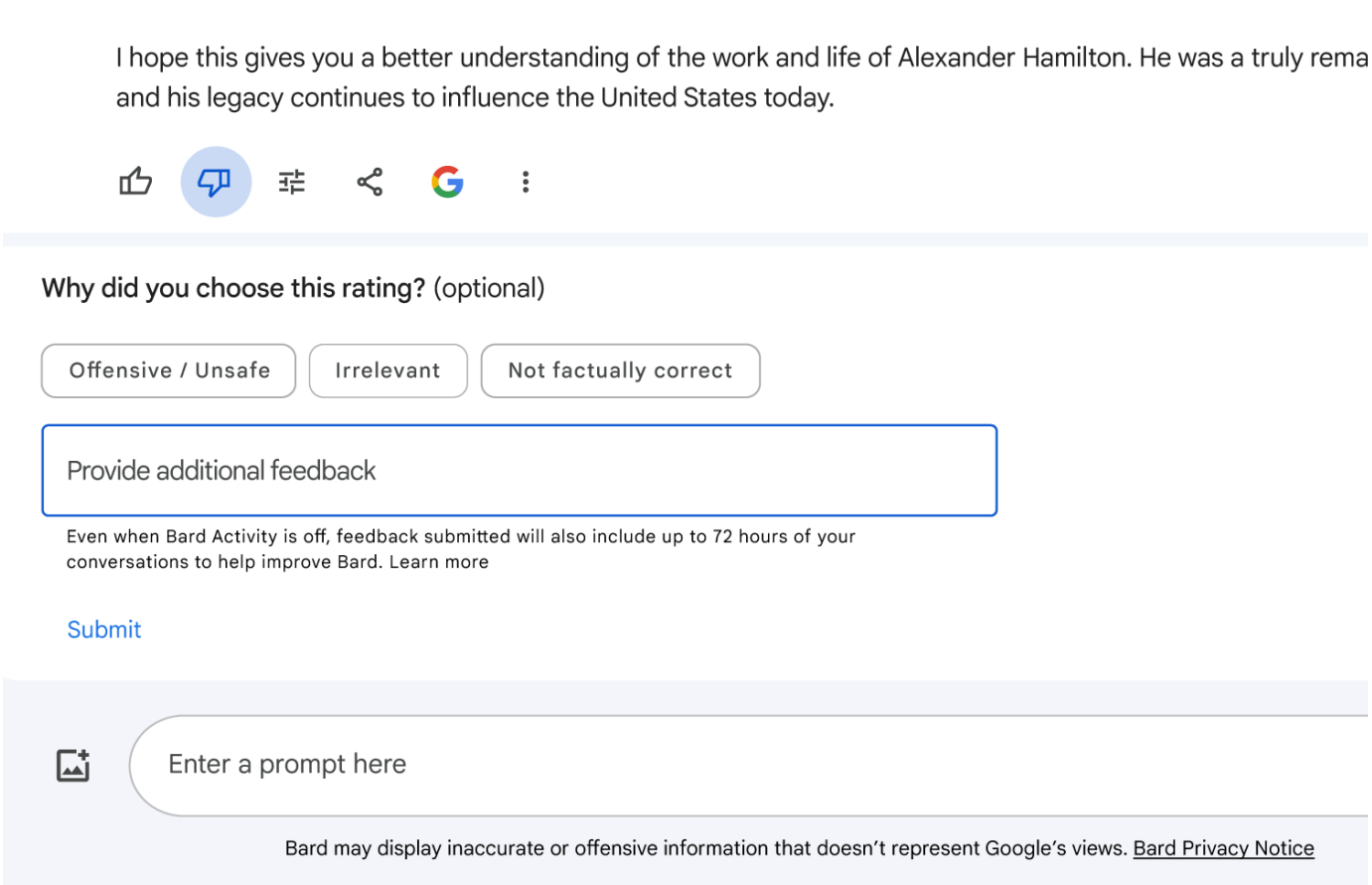
  Conduct adversarial testing to inform Bard's ongoing expansion and feature releases with hundreds to identify potential fairness and safety issues before launches.

- **AI Principle # 4 (Be accountable to people)**

  Provide clear explanations to people to so that they are aware they're interacting with a generative model, and remind people to check Bard's responses. Provide people with information about how Bard works on the landing page.

- **AI Principle # 1 (Be socially beneficial), AI Principle # 2 (Avoid creating or reinforcing unfair bias), AI Principle # 3 (Be built and tested for safety), & AI Principle # 4 (Be accountable to people)**

  Launch as an experiment at first, allow user feedback through "thumbs up" and "thumbs down" buttons, and incorporate that feedback to make improvements to the model.



Bard allows people to provide feedback on its responses with the "thumbs up" and "thumbs down" buttons.

Google continues to explore ways to responsibly develop generative models and the products powered by them. We affirmed our approach to responsible generative AI development in July 2023, when Google joined other industry leaders at the White House to commit to advancing responsible practices in the development of artificial intelligence.

Key Takeaways

Google has company-wide responsible innovation practices that draw upon more than 20 years of work in machine learning (ML) and over a decade of AI research. These practices support developers to incorporate fairness, safety, privacy, and transparency early in development. In addition, Google continues to:

- Prioritize research on societal risks that AI systems can pose, including avoiding harmful bias and discrimination, and protecting individual privacy through transparency and control of personal data.
- Comply with government initiatives, including the White House's industry commitments to ensure safe, secure, and trustworthy AI.
- Work with organizations like the National Institute of Science and Technology in addition to supporting forums such as the Ethical Considerations in Creative Applications of Computer Vision.
- Publish its AI Principles Progress Update report annually.

Bard, an experimental generative AI experience:

- Was Google's first publicly available product powered by a large language model, and, as such, helped establish many precedents for the responsible development of AI.
- Aims to prioritize explainability, fairness, privacy, and safety, with features like "Google it," simple explanations for people, and privacy controls.
- Will continue to evolve in alignment with Google's AI Principles.

Learn more:
g.co/AI/ResponsiblePractices

AI Principles

Since 2018, Google has used these AI Principles to guide the ethical development and use of technology:

- Be socially beneficial.
- Avoid creating or reinforcing unfair bias.
- Be built and tested for safety.
- Be accountable to people.
- Incorporate privacy design principles.
- Uphold high standards of scientific excellence.
- Be made available for use in accord with these principles.

In addition to our principles, Google will not design or deploy AI in the following application areas:

- Those likely to cause overall harm.
- Technologies primarily intended to cause injury.
- Surveillance violating internationally accepted norms.
- Purpose contravenes international law and human rights.

Google's Responsible Innovation Team, which produced this case study, is the company's central AI ethics governance team. It's composed of people with backgrounds in ethics, law, philosophy, research, and various social sciences such as linguistics, economics, political science, international studies, and religious studies.

Internally, we support Googlers in applying the seven principles through AI Principles reviews, education programs, workshops, and other engagements with product teams.

Additional details are available at https://ai.google/principles.

© 2023 Google LLC. All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.