Microsoft

The new Bing: Our approach to Responsible AI

Last updated April 2023

Microsoft

Contents

Microsoft

The basics of the new Bing

Introduction

In February 2023, Microsoft launched the new Bing, an AI-enhanced web search experience. It supports users by summarizing web search results and providing a chat experience. Users can also generate creative content, such as poems, jokes, and letters. The new AI-enhanced Bing runs on a variety of advanced technologies from Microsoft and OpenAI, including GPT-4, a cutting-edge large language model (LLM) from OpenAI. We worked with GPT-4 for months prior to its public release in March 2023 to develop a customized set of capabilities and techniques to join this cutting-edge AI technology and web search.

At Microsoft, we take our commitment to responsible AI seriously. The new Bing experience has been developed in line with Microsoft's AI Principles, Microsoft's Responsible AI Standard, and in partnership with responsible AI experts across the company, including Microsoft's Office of Responsible AI, our engineering teams, Microsoft Research, and Aether. You can learn more about responsible AI at Microsoft here.

In this document, we describe our approach to responsible AI for the new Bing. Ahead of release, we have adopted state-of-the-art methods to identify, measure, and mitigate potential harms and misuse of the system and to secure its benefits for users. As the new Bing becomes available to more users, we know we will continue to learn and evolve our approach. This document will be updated periodically to communicate our evolving processes and methods.

Key terms

The new Bing is an AI-enhanced web search experience. As this is a powerful, new technology, we start by defining some key terms.

Term	Definition
Classifiers	Machine learning models that help to sort data into labeled classes or categories of information. In the new Bing, one way in which we use classifiers is to help detect potentially harmful content submitted by users or generated by the system in order to mitigate generation of that content and misuse or abuse of the system.
Grounding, Grounded responses	The new Bing is grounded in web search results when users are seeking information. This means that we center the response provided to a user's query or prompt on high-ranking content from the web, and we provide links to websites so that users can learn more. Bing ranks web search content by heavily weighting features such as relevance, quality and credibility, and freshness. We describe these concepts in more detail in Bing's Webmaster Guidelines (see "Quality and Credibility" in "How Bing ranks your content"). We consider **grounded responses** to be responses from the new Bing in which statements are supported by information contained in input sources, such as web search results from the query or prompt, Bing's knowledge

	base of fact-checked information, and, for the chat experience, recent conversational history from the chat. Ungrounded responses are those in which a statement is not grounded in those input sources.
Large language models (LLMs)	Large language models (LLMs) in this context are AI models that are trained on large amounts of text data to predict words in sequences. LLMs are capable of performing a variety of language tasks, such as text generation, summarization, translation, classification, and more.
Metaprompt	The metaprompt is a program that serves to guide the system's behavior. Parts of the metaprompt help align system behavior with Microsoft AI Principles and user expectations. For example, the metaprompt may include a line such as "communicate in the user's language of choice."
Mitigation	A method or combination of methods designed to reduce potential harms that may arise from using AI-driven features in the new Bing.
Prompt	The text a user sends to Bing from the search bar or in the chat experience. The prompt is used as an input to the model that powers the new Bing in the chat experience. For example, a user might input the following prompt: "*I am planning a trip for our anniversary in September. What are some places we can go that are within a 3-hour flight from London Heathrow?*"
Query	The text a user sends to Bing from the search bar for web search. In this document, we distinguish a query (for web search) from a prompt (for the chat experience). For example, a user might input the following query to search the web: "*travel destinations near London Heathrow airport*"
Red team testing and stress-testing	Techniques used by experts to assess the limitations and vulnerabilities of a system and to test the effectiveness of planned mitigations. Red team testing and stress-testing are used to identify potential harms and are distinct from systematic measurement of harms.
Response	The text that new Bing outputs in response to a prompt. Synonyms for "response" include "completion," "generation," and "answer." For example, Bing may provide the following response to the prompt that we used in the definition of 'prompt above: "*Congratulations on your anniversary! 🎉 There are many places you can go that are within a 3-hour flight from London Heathrow. According to one source, some of the weekend escapes within three hours by plane from London are Bilbao, Copenhagen, Reykjavik, and Stockholm.*[1,2] *Another source*

suggests some of the best places to go in Europe in September are Krakow, Florence, and Lisbon.[3,4] You can also check the direct flights from London Heathrow to various destinations and see what appeals to you.[5] I hope you have a wonderful trip!"

Intended uses and new AI experiences

The new Bing's intended uses are to connect users with relevant search results, review results from across the web to find and summarize answers users are looking for, help users refine their research to get answers with a chat experience, and spark creativity by helping users create content. The new Bing's generative AI experiences below support the goal of being an AI-powered copilot for the web.

- **Summarization**. When users submit a search query on the new Bing, the Bing system processes the query, conducts one or more web searches, and uses the top web search results to generate a summary of the information to present to users. These summaries include references to help users see and easily access the search results used to help ground the summary. Summaries can appear on the right side of the search results page and within the chat experience.
- **Chat experience**. In addition to summarization, users can chat with the new Bing system via text or voice input, ask follow-up questions to clarify searches and find new information, and submit prompts to generate creative content. References are also included in the chat experience when the new Bing is summarizing search results in the response.
- **Generation of creative content**. In both the chat experience and on the search page, users can create poems, jokes, stories, letters, and other content with help from the new Bing.

How does the new Bing work?

With the new Bing, we've developed an innovative approach to bring state-of-the-art LLMs to web search. When a user enters a prompt in the new Bing, the prompt, recent conversation history, the metaprompt, and top search results are sent as inputs to the LLM. The model generates a response using the user's prompt and recent conversation history to contextualize the request, the metaprompt to align responses with Microsoft AI Principles and user expectations, and the search results to ground responses in existing, high-ranking content from the web.

Responses are presented to users in several different formats, such as traditional links to web content, AI-generated summarizations, and chat responses. Summarizations and chat responses that rely on web search results will include references and a "Learn more" section below the responses, with links to search results that were used to ground the response. Users can click these links to learn more about a topic and the information used to ground the summary or chat response.

In the chat experience, users can perform web searches conversationally by adding context to their prompt and interacting with the system responses to further specify their search interests. For example, a user might ask follow-up questions, request additional clarifying information, or respond to the system in a conversational way. In the chat experience, users can also select a response from pre-written suggestions, which we call chat suggestions. These buttons appear after each response from Bing and provide suggested prompts to continue the conversation within the chat experience. Chat suggestions also appear alongside summarized content on the search results page as an entry point for the chat experience.

On both the search results page and in the chat, the creator experience allows a user to create stories, poems, song lyrics, and letters with help from Bing. When Bing detects user intent to generate creative content (for example, the prompt begins with "write me a …"), the system will, in most cases, generate content responsive to the user's prompt.

Bing strives to provide diverse and comprehensive search results with its commitment to free and open access to information. At the same time, Bing's product quality efforts include working to avoid inadvertently promoting potentially harmful content to users. More information on how Bing ranks content, including how it defines relevance, quality, and credibility of a webpage, is available in the "Bing Webmaster Guidelines." More information on Bing's content moderation principles is available in "How Bing delivers search results."

Identifying, measuring, and mitigating harms

Like other transformational technologies, harnessing the benefits of AI is not risk-free, and a core part of Microsoft's Responsible AI program is designed to identify potential harms, measure their propensity to occur, and build mitigations to address them. Guided by our AI Principles and our Responsible AI Standard, we sought to identify, measure, and mitigate potential harms and misuse of the new Bing while securing the transformative and beneficial uses that the new experience provides. In the sections below we describe our iterative approach to identify, measure, and mitigate potential harms.

Identify

At the model level, our work began with exploratory analyses of GPT-4 in the late summer of 2022. This included conducting extensive red team testing in partnership with OpenAI. This testing was designed to assess how the latest technology would work without any additional safeguards applied to it. Our specific intention at this time was to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Our combined learnings across OpenAI and Microsoft contributed to advances in model development and, for us at Microsoft, informed our understanding of risks and contributed to early mitigation strategies for the new Bing.

In addition to model-level red team testing, a multidisciplinary team of experts conducted numerous rounds of application-level red team testing on the new Bing AI experiences before making them publicly available in our limited release preview. This process helped us better understand how the system could be exploited by adversarial actors and improve our mitigations. Non-adversarial stress-testers also extensively evaluated new Bing features for shortcomings and vulnerabilities. Post-release, the new AI experiences in Bing are integrated into the Bing engineering organization's existing production measurement and testing infrastructure. For example, red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Bing uses for improving the system.

Measure

Red team testing and stress-testing can surface instances of specific harms, but in production users will have millions of different kinds of conversations with the new Bing. Moreover, conversations are multi-turn and contextual, and identifying harmful content within a conversation is a complex task. To better understand and address the potential for harms in the new Bing AI experiences, we developed additional RAI metrics specific to those new AI experiences for measuring potential harms like jailbreaks, harmful content, and ungrounded content. We also enabled measurement at scale through partially automated

measurement pipelines. Each time the product changes, existing mitigations are updated, or new mitigations are proposed, we update our measurement pipelines to assess both product performance and the new RAI metrics.

As an illustrative example, the updated partially automated measurement pipeline for harmful content includes two major innovations: conversation simulation and automated, human-verified conversation annotation. First, RAI experts built templates to capture the structure and content of conversations that could result in different types of harmful content. These templates were then given to a conversational agent which interacted as a hypothetical user with the new Bing, generating simulated conversations. To identify whether these simulated conversations contained harmful content, we took guidelines that are typically used by expert linguists to label data and modified them for use by GPT-4 to label conversations at scale, refining the guidelines until there was significant agreement between model-labeled conversations and human-labeled conversations. Finally, we used the model-labeled conversations to calculate an RAI metric that captures the effectiveness of the new Bing at mitigating harmful content.

Our measurement pipelines enable us to rapidly perform measurement for potential harms at scale. As we identify new issues through the preview period and ongoing red team testing, we continue to expand the measurement sets to assess additional harms.

Mitigate

As we identified potential harms and misuse through processes like red team testing and stress-testing and measured them with the innovative approaches described above, we developed additional mitigations to those used for traditional search. Below, we describe some of those mitigations. We will continue monitoring the new Bing AI experiences to improve product performance and mitigations.

Preview period, phased release via waitlist. We are committed to learning and improving our responsible AI approach continuously as our technologies and user behavior evolve. Our incremental release strategy has been a core part of how we move our technology safely from the labs into the world, and we're committed to a deliberate, thoughtful process to secure the benefits of the new Bing. Limiting the number of people with access during the preview period allows us to discover how people use the new Bing, including how people may misuse it, so we can try to mitigate emerging issues before broader release. We are making changes to the new Bing daily to improve product performance, improve existing mitigations, and implement new mitigations in response to our learnings during the preview period.

Account authentication. We are requiring users to authenticate using their Microsoft account before using the full new Bing experience. This discourages abuse and helps us to take appropriate action in response to Code of Conduct violations, when needed.

Grounding in search results. As noted above, the new Bing is designed to provide responses supported by the information in web search results when users are seeking information. For example, the system is provided with text from the top search results and instructions via the metaprompt to ground its response. However, in summarizing content from the web, the new Bing may include information in its response that is not present in its input sources. In other words, it may produce ungrounded results. Our early evaluations have indicated that ungrounded results in chat may be more prevalent for certain types of prompts or topics than others, such as asking for mathematical calculations, financial or market information (for example, company earnings, stock performance data), and information like precise dates of events or specific prices of items. Users should always take caution and use their best judgement when viewing summarized search results, whether on the search results page or in the chat experience. We have

taken several measures to mitigate the risk that users may overrely on ungrounded generated content in summarization scenarios and chat experiences. For example, responses in the new Bing that are based on search results include references to the source websites for users to verify the response and learn more. Users are also provided with explicit notice that that they are interacting with an AI system and advised to check the web results source materials to help them use their best judgement. We will continue to explore additional mitigation approaches for ungrounded content during and after our preview release period.

AI-based classifiers and metaprompting to mitigate harms or misuse. The use of LLMs may produce problematic content that could lead to harms or misuse. Examples could include output related to self-harm and violence, graphic content, intellectual property, inaccurate information, hateful speech, or text that could relate to illegal activities. Classifiers and metaprompting are two examples of mitigations that have been implemented in the new Bing to help reduce the risk of these types of content. ***Classifiers*** classify text to flag different types of potentially harmful content in search queries, chat prompts, or generated responses. Bing uses AI-based classifiers and content filters, which apply to all search results and relevant features; we designed additional prompt classifiers and content filters specifically to address possible harms raised by the new Bing features such as chat. Flags lead to potential mitigations, such as not returning generated content to the user, diverting the user to a different topic, or redirecting the user to traditional search. ***Metaprompting*** involves giving instructions to the model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. For example, the metaprompt may include a line such as "communicate in the user's language of choice."

Limiting conversational drift. During the preview period we have learned that very long chat sessions can result in responses that are repetitive, unhelpful, or inconsistent with new Bing's intended tone. To address this conversational drift, we have limited the number of turns (exchanges which contain both a user question and a reply from Bing) per chat session. We continue to evaluate additional approaches to mitigate this issue.

User-centered design and user experience interventions. User-centered design and user experiences are an essential aspect of Microsoft's approach to responsible AI. The goal is to root product design in the needs and expectations of users. As users interact with the new Bing for the first time, we offer various touchpoints designed to help them understand the capabilities of the system, disclose to them that the new Bing is powered by AI, and communicate limitations. The experience is designed in this way to help users get the most out of the new Bing and minimize the risk of overreliance. Elements of the experience also help users better understand the new Bing and their interactions with it. These include chat suggestions specific to responsible AI (for example, How does Bing use AI? Why won't Bing respond on some topics?), explanations of limitations, ways users can learn more about how the system works and report feedback, and easily navigable references that appear in responses to show users the results and pages in which responses are grounded.

AI disclosure. The new Bing provides several touchpoints for meaningful AI disclosure where users are notified that they are interacting with an AI system as well as opportunities to learn more about the new Bing. Empowering users with this knowledge can help them avoid over-relying on AI and learn about the system's strengths and limitations.

Terms of Use, including Code of Conduct. This resource governs use of the new Bing. Users have easy access to the Terms of Use and Code of Conduct, which, among other things, inform them of permissible and impermissible uses and the consequences of violating terms. The Terms of Use also provides additional disclosures for users and serves as a handy reference for users to learn about the new Bing.

Operations and incident response. We also use Bing's ongoing monitoring and operational processes to address when the new Bing receives signals, or receives a report, indicating possible misuse or violations of the Terms of Use or Code of Conduct.

Feedback, monitoring, and oversight. The new Bing experience builds on existing tooling that allows users to submit feedback and report concerns, which are reviewed by Microsoft's operations teams. Bing's operational processes have also expanded to accommodate the features within the new Bing experience, for example, updating the Report a Concern page to include the new types of content that users generate with the help of the model.

Our approach to identifying, measuring, and mitigating harms will continue to evolve as we learn more and we are already making improvements based on feedback gathered during the preview period.

Protecting young users

Microsoft considers the best interest of its users as core to product design. Presently, child account holders cannot sign up for the new Bing while we continue to research how children of different age groups will experience the new Bing features. We will continue to assess how the new Bing can provide opportunities to augment and improve young users' well-being and how we can appropriately mitigate potential harms or risks.

As described above, we have implemented safeguards that mitigate potentially harmful content and we provide parental controls to enable appropriate use by younger users of the new Bing. We have designed Bing's new features to limit the production of potentially offensive, harmful, or illegal materials that could negatively affect users, including young users. In addition to information we have provided in this document and in our FAQs regarding chat features, more information about how the new Bing works to avoid responding with unexpected offensive content in search results is available here. Family Safety settings provide parents and guardians the option to engage additional protections, such as controlling the level of Safe Search protection for their family users as well as seeing search and browse history of those users within their family accounts.

Finally, Microsoft has committed to not deliver personalized advertising based on online behavior to children whose birthdate in their Microsoft account identifies them as under 18 years of age. This important protection extends to ads in the new Bing features. Users may see contextual ads based on the query or prompt used to interact with Bing.

Protecting privacy

Microsoft's longstanding belief that privacy is a fundamental human right has informed every stage of Microsoft's development and deployment of the new Bing experience. Our commitments to protecting the privacy of all users, including by providing individuals with transparency and control over their data and integrating privacy by design through data minimization and purpose limitation, are foundational to the new Bing. As we evolve our approach to providing the new Bing's generative AI experience, we will continually explore how best to protect privacy. This document will be updated as we do so. More information about how Microsoft protects our users' privacy is available in the Microsoft Privacy Statement.

Protecting children and young people

As described above, access to the new Bing is disallowed to all Microsoft accounts that require parental consent under local laws, for example users under 13 in the U.S. Protections available for Microsoft's teen users through the Microsoft account extend to their use of the new Bing features. For example, parents can opt to engage additional protections through Family Safety settings, such as controlling the level of Safe Search protection, as well as seeing search and browse history of teen users within their family accounts. Our existing commitment not to deliver personalized advertising to children whose birthdate in their Microsoft account identifies them as under 18 years of age extends to the new Bing.

Transparency and control

To unlock the transformative potential of generative AI, we must build trust in the technology through empowering individuals to understand how their data is used and providing them with meaningful choices and controls over their data. The new Bing is designed to prioritize human agency, through providing information on how the product works as well as its limitations, and through extending our robust consumer choices and controls to the new Bing features.

The Microsoft Privacy Statement provides information about our transparent privacy practices for protecting our customers, and it sets out information on the controls that give our users the ability to view and manage their personal data. To help ensure that users have the information they need when they are interacting with Bing's new conversational features, in-product disclosures inform users that they are engaging with an AI product, and we provide links to further FAQs and explanations about how these features work. Microsoft will continue to listen to user feedback and will add further detail on Bing's conversational features as appropriate to support understanding of the way the product works.

Microsoft also provides its users with robust tools to exercise their rights over their personal data. For data that is collected by the new Bing, including through user queries and prompts, the Microsoft privacy dashboard provides users with tools to exercise their data subject rights, including by providing users with the ability to view, export, and delete stored conversation history.

The new Bing also honors requests under the European right to be forgotten, following the process that Microsoft developed and refined for Bing's traditional search functionality. All users can report concerns regarding generated content and responses here, and our European users can use this form to submit requests to block search results in Europe under the right to be forgotten.

The new Bing will honor users' privacy choices, including those that have previously been made, such as consent for data collection and use that is requested through cookie banners. To help enable user autonomy and agency in making informed decisions, we have used our internal review process to carefully examine how choices are presented to users.

Privacy by design through purpose limitation and data minimization

The new Bing was built with privacy in mind, so that personal data is collected and used only as needed and is retained no longer than is necessary. More information about the personal data that Bing collects, how it is used, and how it is stored and deleted is available in the Microsoft Privacy Statement, which also provides information about Bing's new chat features. The data collected by the new Bing is used only to provide the service and contextually relevant advertisements.

The new Bing has data retention and deletion policies to help ensure that personal data collected through Bing's chat features is only kept as long as needed.

We will continue to learn and evolve our approach in providing the new Bing, and as we do so we will continue to work across disciplines to align our AI innovation with human values and fundamental rights, including protecting young users and privacy.

Learn more about responsible AI

This document is part of a broader effort at Microsoft to put our AI principles into practice. To find out more, see Microsoft's Approach to Responsible AI.

Microsoft's Responsible AI Standard

Microsoft's responsible AI resources

Microsoft Azure Learning courses on responsible AI

Learn more about the new Bing

Introducing the new Bing

Terms of Use and Code of Conduct

About this document