Microsoft

2025 Responsible AI Transparency Report

How we build,
support our customers,
and grow

Contents

Foreword

In May 2024, we released our inaugural Responsible AI Transparency Report. We're grateful for the feedback we received from our stakeholders around the world. Their insights have informed this second annual Responsible AI Transparency Report, which underscores our continued commitment to building AI technologies that people trust. Our report highlights new developments related to how we build and deploy AI systems responsibly, how we support our customers and the broader ecosystem, and how we learn and evolve.

The past year has seen a wave of AI adoption by organizations of all sizes, prompting a renewed focus on effective AI governance in practice. Our customers and partners are eager to learn about how we have scaled our program at Microsoft and developed tools and practices that operationalize high-level norms. Like us, they have found that building trustworthy AI is good for business, and that good governance unlocks AI opportunities. According to IDC's Microsoft Responsible AI Survey that gathered insights on organizational attitudes and the state of responsible AI, over 30% of the respondents note the lack of governance and risk management solutions as the top barrier to adopting and scaling AI.[1] Conversely, more than 75% of the respondents who use responsible AI tools for risk management say that they have helped with data privacy, customer experience, confident business decisions, brand reputation, and trust.

We've also seen new regulatory efforts and laws emerge over the past year. Because we've invested in operationalizing responsible AI practices at Microsoft for close to a decade, we're well prepared to comply with these regulations and to empower our customers to do the same. Our work here is not done, however. As we detail in the report, efficient and effective regulation and implementation practices that support the adoption of AI technology across borders are still being defined. We remain focused on contributing our practical insights to standard- and norm-setting efforts around the world.

Across all these facets of governance, it's important to remain nimble in our approach, applying learnings from our real-world deployments, updating our practices to reflect advances in the state-of-the-art, and ensuring that we are responsive to feedback from our stakeholders. Learnings from our principled and iterative approach are reflected in the pages of this report. As our governance practices continue to evolve, we'll proactively share our fresh insights with our stakeholders, both in future annual transparency reports and other public settings.

In the year ahead, we will focus on developing more flexible and agile risk management techniques, advancing a vibrant ecosystem through shared norms and effective tools, and supporting effective governance across the AI supply chain. Together, these efforts will enhance the abilities of not only AI developers and deployers like Microsoft, but also our customers and partners, to implement governance efficiently and effectively, fostering the trust we need at a pace that matches AI innovation.

We look forward to continuing to earn, build, and keep trust in AI technology to help people around the world benefit from its profound potential.

Teresa J Hutson

Teresa Hutson
Corporate Vice President

Natasha Crampton

Natasha Crampton
Chief Responsible AI Officer

Key takeaways

In 2024, we made key investments in our responsible AI tools, policies, and practices to move at the speed of AI innovation.

1 **We improved our responsible AI tooling** to provide expanded risk measurement and mitigation coverage for modalities beyond text—like images, audio, and video—and additional support for agentic systems, semi-autonomous systems that we anticipate will represent a significant area of AI investment and innovation in 2025 and beyond.

2 **We took a proactive, layered approach to compliance** with new regulatory requirements, including the European Union’s AI Act, and provided our customers with resources and materials that empower them to innovate in line with relevant regulations. Our early investments in building a comprehensive and industry-leading responsible AI program positioned us well to shift our AI regulatory readiness efforts into high gear in 2024.

3 **We continued to apply a consistent risk management approach** across releases through our pre-deployment review and red teaming efforts. This included oversight and review of high-impact and higher-risk uses of AI and generative AI releases, including every flagship model added to the Azure OpenAI Service and every Phi model release. To further support responsible AI documentation as part of these reviews, we launched an internal workflow tool designed to centralize the various responsible AI requirements outlined in the Responsible AI Standard.

4 **We continued to provide hands-on counseling for high-impact and higher-risk uses of AI** through our Sensitive Uses and Emerging Technologies team. Generative AI applications, especially in fields like healthcare and the sciences, were notable growth areas in 2024. By gleaning insights across cases and engaging researchers, the team provided early guidance for novel risks and emerging AI capabilities, enabling innovation and incubating new internal policies and guidelines.

5 **We continued to lean on insights from research** to inform our understanding of sociotechnical issues related to the latest advancements in AI. We established the AI Frontiers lab to invest in the core technologies that push the frontier of what AI systems can do in terms of capability, efficiency, and safety.

6 **We worked with stakeholders around the world to make progress** towards building coherent governance approaches to help accelerate adoption and allow organizations of all kinds to innovate and use AI across borders. This included publishing a book exploring governance across various domains and helping advance cohesive standards for testing AI systems.

Section 1

How we build generative AI systems and models responsibly

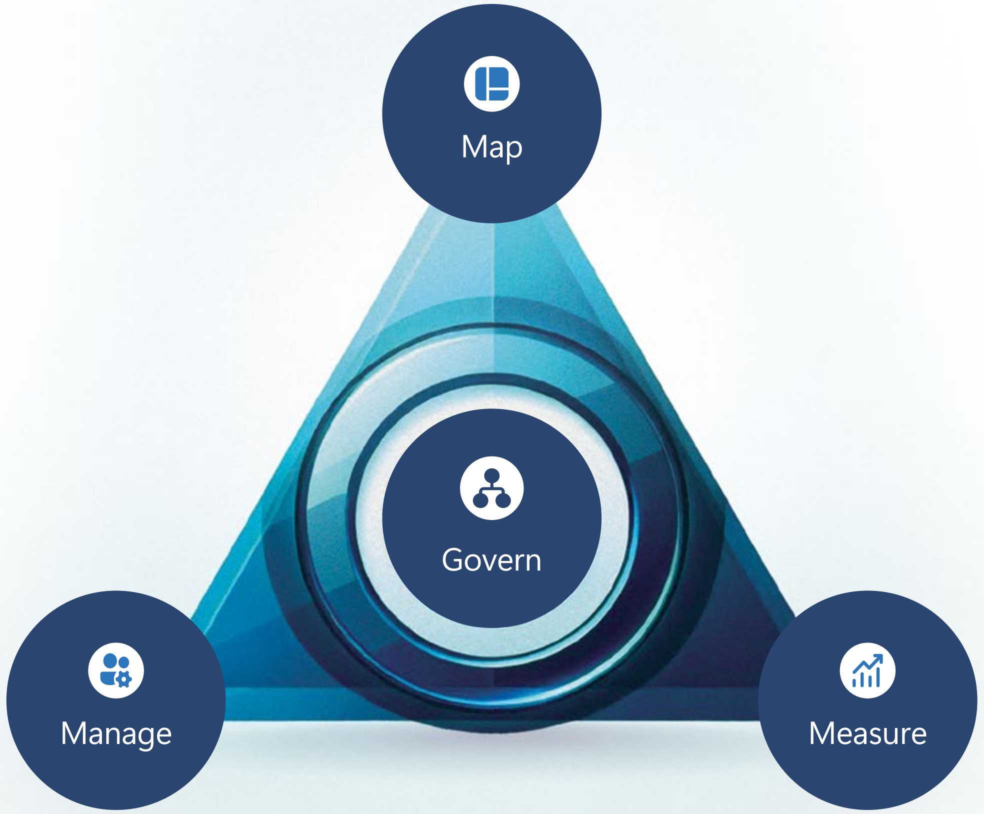
At Microsoft, we remain focused on our mission of empowering every person and organization to achieve more. This remains true in the age of AI, where the potential for accelerated human achievement is greater than ever. We are clear-eyed about the role we play in shaping this technology and in our understanding that people do not use technology they do not trust. For us, the daily work of earning trust in the age of AI requires keeping humans at the center of how we design, develop, deploy AI—a practice that started in 2016 with the first draft of our AI principles.[2]

Formally adopted in 2018, our AI principles of fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability continue to serve as our enduring North Star. Over the years, we have continually referred back to these principles when new AI technologies unlock previously unforeseen capabilities and risks. Our principles provide us with the contours for establishing new policies, tools, and practices, or refining existing ones, as AI technology and regulation continue to rapidly evolve

When we develop and deploy a new generative AI systems and models, we leverage the AI Risk Management Framework created by the National Institute for Standards and Technology (NIST),[3] which includes four key functions: govern, map, measure, and manage. In this section, we describe how these four functions guide how we develop and deploy AI, highlighting the changes and progress we have made since the publication of our first transparency report in May 2024.

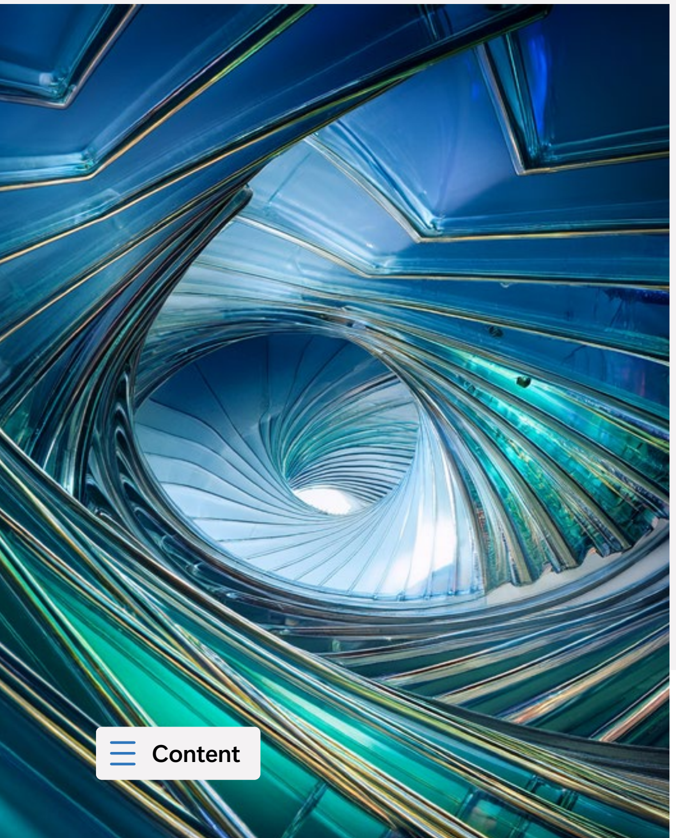
How we build generative AI systems

We use a multi-layered approach to manage and mitigate risks for generative AI systems and models across the AI development lifecycle.



Govern: Policies, practices, and processes

Our responsible AI governance architecture is designed to help us uphold our AI principles consistently across the AI development lifecycle. Our governance work defines policies that align with our principles, articulates roles and responsibilities for carrying out these policies, enables teams throughout Microsoft to foster a culture of proactive risk management, and establishes processes that reinforce a responsible AI by design approach.[4]



Content

Policies and practices

The bedrock of our governance work, and our AI work writ large, is our Responsible AI Standard. The Standard, which we first developed in 2019 and later revamped and released publicly in 2022, serves as our internal playbook for building AI systems in alignment with our six AI principles.[5] As new AI capabilities, risks, and regulatory requirements emerge, we build on the Standard to refine existing requirements or define new ones.

In 2023, we formalized a set of specific internal requirements for generative AI systems to help us navigate the novel risks they presented. In 2024, we continued to update and improve these requirements, including establishing new policies for model development and deployment as part of our proactive, layered approach to compliance with new regulatory requirements, including the European Union's AI Act. Cross-functional working groups identified key requirements to help our Microsoft teams get ready for enforcement deadlines and to support our customers with their own compliance efforts.

Read more about the EU AI Act on page 10 ›

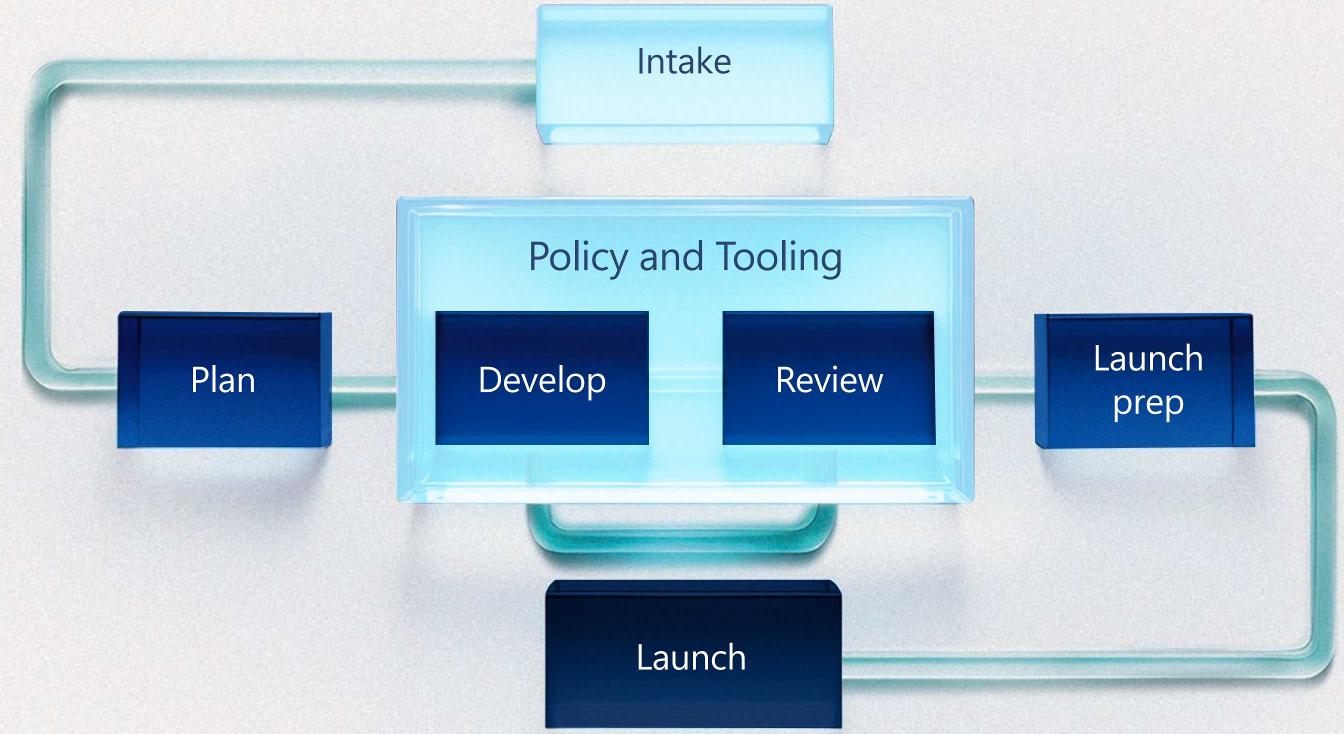
Our new policies for model development and deployment include our Frontier Governance Framework, which we shared with the public in February 2025.[6] This framework originated from the voluntary Frontier AI Safety Commitments that Microsoft and fifteen other AI organizations made in May 2024 with the support of governments from around the world. The framework serves as a monitoring function, tracking the emergence of new and advanced AI model capabilities that could be misused to threaten national security or pose at-scale public safety risks. It also sets out a process for assessing and mitigating these risks so that frontier AI models can be deployed in a secure and trustworthy way.

Our Frontier Governance Framework integrates with our broader AI governance program by drawing on best practices for risk assessment, testing, and safety and security mitigations. We expect this framework to be updated over time as our understanding of AI risk and risk mitigation techniques improves. We look forward to working with others across the industry, government, and civil society to develop and apply best practices through this framework.

Given the rapid speed of AI innovation and regulatory developments around the world, we focused on streamlining our policy-to-implementation pipeline in 2024. The first step in this pipeline consists of our Office of Responsible AI identifying and prioritizing the development of new policies or guidance. This step is informed by regulatory developments around the world; pre-deployment review processes, red teaming operations, and product roadmaps; and other signals from ongoing monitoring and incident response efforts.

From there, the Office of Responsible AI leads a policy development process that involves consultation with a multi-stakeholder group of Microsoft researchers, engineering teams, legal counsel, and other policy experts. In parallel to policy development, the Responsible AI engineering team develops tools and specific technical instructions for teams. Finally, the new policy and accompanying resources are built together to be socialized amongst engineering teams before launch to help them understand expectations and available resources.

Responsible AI Policy to Implementation Pipeline



EU AI Act implementation efforts

Abiding by the laws of every jurisdiction in which we operate is core to our business practice. Our customers look to us to build technology that is trustworthy, which means keeping pace with rapidly evolving regulations and technological developments and communicating updates to our practices and policies as we do so.

A prime example of this is our implementation of the EU AI Act, a comprehensive new law that establishes uniform rules across all EU Member States to address potential AI risks to health, safety, and fundamental rights. The EU intentionally phased the Act's implementation over several years to ensure companies have adequate time to meet the Act's regulatory requirements, with initial obligations going into effect on February 2, 2025.

The AI Act spreads obligations across actors in the AI supply chain, with different obligations applying to different entities, including providers, deployers, distributors, and importers. The Act's obligations are tied to development and deployment activities in the European Union, so its provisions apply to companies that operate in non-EU countries, including the United States, when they place AI products on the EU market or produce AI systems whose outputs are used in the EU.[7]

Microsoft's early investments in building a comprehensive and industry-leading responsible AI program positioned us well to do two things:

1. Expand our already accelerated AI regulatory readiness efforts into broader Microsoft, customer, and partner safety and security investments
2. Share our insights from years of developing and implementing AI policies and working with AI regulators around the world, including the EU

Within Microsoft, cross-functional working groups combining AI governance, engineering, legal, and public policy experts worked together to identify whether and how our internal standards and practices should be updated to reflect the final text of the AI Act and details emerging from the General Purpose AI (GPAI) Code of Practice. The first set of provisions that came into effect in February 2025 were the AI literacy requirements and the prohibited practices provisions that ban the use of AI for practices such as social scoring and real-time biometric identification. Microsoft has undertaken various initiatives to promote AI literacy in accordance with the EU AI Act, empowering our employees, customers, and others to responsibly leverage AI technologies.

Read more about our AI literacy efforts on page 49 ›

In alignment with these compliance deadlines, we proactively took a layered approach to prepare for compliance with the prohibited practices provisions, including:

- **Conducting a thorough review of Microsoft-owned systems** already on the market to identify any places where we needed to adjust our approach, including by updating documentation or implementing technical mitigations. To do this, we developed a series of questions designed to elicit whether an AI system could implicate a prohibited practice and dispatched this survey to our engineering teams via our central tooling. Relevant experts reviewed the responses and followed up with teams directly where further clarity or additional steps were necessary. These screening questions remain in our central responsible AI workflow tool so that teams working on new AI systems can answer them and engage the review workflow as needed.
- **Creating new restricted uses** internally and updating our policies to ensure Microsoft does not design or deploy AI systems for uses prohibited by the EU AI Act. We also developed specific marketing and sales guidance to ensure that our general-purpose AI technologies are not marketed or sold for uses that could implicate the EU AI Act's prohibited practices.
- **Updating our contracts**, including our Microsoft Enterprise AI Services Code of Conduct,[8] so that our customers clearly understand they cannot engage in any prohibited practices. For example, the Microsoft Enterprise AI Services Code of Conduct now has an express prohibition on the use of the services for social scoring.

In 2024, Microsoft joined a multistakeholder process to support the development of the EU AI Act Code of Practice for GPAI models. The Code of Practice for GPAI models is intended to include a set of guidelines for compliance with the AI Act's GPAI model provider obligations, which come into effect in August 2025. Taking early signal from the set of guidelines emerging from the Code of Practice, we took a similar layered approach to prepare for compliance. This included:

- **Creating new model-level policy requirements** to help our product teams understand the set of evaluations, documentation, and disclosure obligations they will need to complete when building models.
- **Building new workflows and automation** to capture the information and generate the documentation needed to align with model policies. For example, we are in the process of integrating our central responsible AI workflow with model training infrastructure to support the automatic capture of model training details for transparency obligations.

Microsoft continues to engage with the central EU regulator, the AI Office, and other relevant authorities in EU Member States to share insights from our AI development, governance, and compliance experience, as well as insights we hear from our customers. We also continue to seek clarity on open questions and advocate for practical outcomes that are efficient, effective, and interoperable internationally.

In the second half of 2025, we plan to enhance our reporting with new information detailing the content used for training GPAI models developed by Microsoft. This aligns with our commitment to transparency, upcoming regulatory requirements, and our support for downstream responsible use of our GPAI models.

Processes

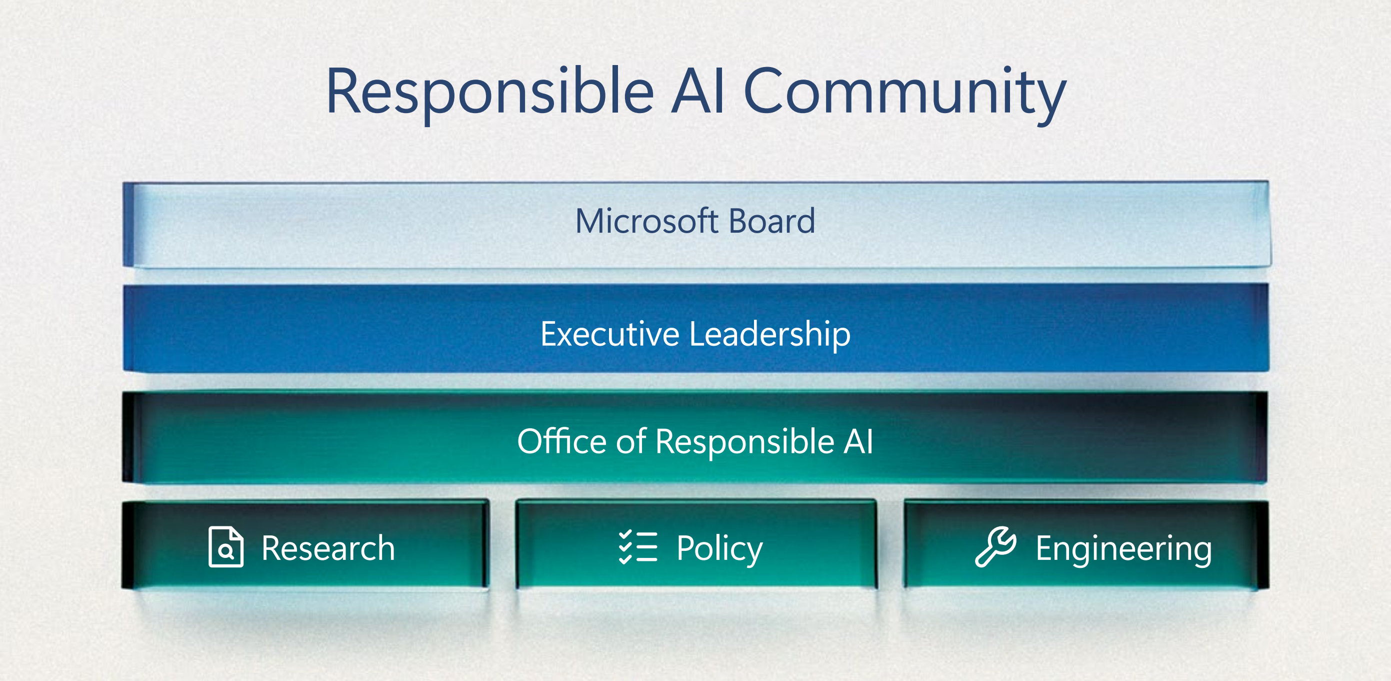
While establishing principles and policies are a critical building block to our responsible AI program, it takes a broad, cross-company effort to bring Microsoft's governance framework to life.

At the executive level, leadership support for our responsible AI program starts with our CEO Satya Nadella and our Responsible AI Council, led by Vice Chair and President Brad Smith and Chief Technology Officer and Executive Vice President of AI, Kevin Scott. Quarterly Responsible AI Council meetings provide executive-level oversight into the company-wide progress we are making on our commitments. Our progress is also regularly reported to, and guidance is solicited from, the Microsoft Board of Directors through the Board's Environmental, Social, and Public Policy (ESPP) committee.

Orchestrating Microsoft's responsible AI program across the company requires the muscle of a broader governance community. At Microsoft, this includes the Office of Responsible AI and a dedicated network of Responsible AI leaders and champions embedded throughout divisions across the company. The Office of Responsible AI advises teams across the company on legal and regulatory requirements and manages the Responsible AI Governance Community by defining roles and responsibilities, creating documented processes, and leading oversight processes such as the Sensitive Uses and Emerging Technologies program.

Over the years, our Responsible AI Governance Community has matured, creating more specialized roles within each division at Microsoft. Since the launch of our Responsible AI Champions program in 2020, our Responsible AI Governance Community has grown to include Responsible AI Corporate Vice Presidents (CVPs),[9] Division Leads, and Lead Responsible AI Champs.

Responsible AI CVPs are accountable executives who provide oversight of their group's implementation of and adherence to responsible AI policies and serve on the Responsible AI Council. They are kept informed of progress by a Division Lead who drives operations and implementation of our processes to uphold customer trust in Microsoft's AI-powered products and services. Division Leads partner with Lead Responsible AI Champs to keep their teams informed of updates, implement procedures, and ensure adherence to our policies.



Responsible AI Community

Microsoft Board

Executive Leadership

Office of Responsible AI

Research

Policy

Engineering

> We continue to invest in growing, training, and cultivating a thriving and empowered Responsible AI community, which includes the Responsible AI Governance Community and other teams like RAI engineering, Aether,[10] Microsoft Research, and the AI Red Team (AIRT) that carry out critical functions in the mapping, measurement, and management of AI risks.

The Responsible AI community is provided with in-depth, ongoing training to equip them to implement responsible AI practices within their teams and divisions. In 2024, Responsible AI community members participated in a variety of trainings that covered responsible AI policies, procedures, and tools. Trainings included updates to responsible AI policies as we prepared to implement new regulatory requirements, new guidance and tooling to support teams in measuring and mitigating risks, and specialized topics at the intersection of AI and security.

Throughout 2024, we continued to offer all Microsoft employees training on responsible AI and AI more broadly that catered to different technical knowledge levels and the contexts in which they develop or use AI. This included both live training and self-paced training. At the broadest level, all Microsoft employees were required to complete Trust Code (Standards of Business Conduct), our companywide ethics course, which included training on responsible AI. As of January 22, 2025, 99 percent of all employees completed this course.

In addition to this training, employees also have the option to participate in hackathons and learning series focused on responsible AI. Throughout 2024, there were seven learning sessions focused on responsible AI hosted as part of the AI/ML Learning Series, which features insights from research and practice on AI. Cumulatively, these responsible AI-focused sessions had 6,798 attendees. In late 2024, a responsible AI-focused hybrid event featured 22 deep-dive sessions on insights and best practices related to responsible AI. This event had 1,020 unique attendees at the live session and the content was posted online for employees across the company to view at their own pace.

For more hands-on learning experience, Microsoft's annual Hackathon offers employees an opportunity to step out of their day-to-day work and team up with colleagues from across the company to build something innovative. The 2024 Hackathon had 738 hacks on AI that included a focus on responsible AI.

99%

of all employees completed the Trust Code (Standards of Business Conduct)

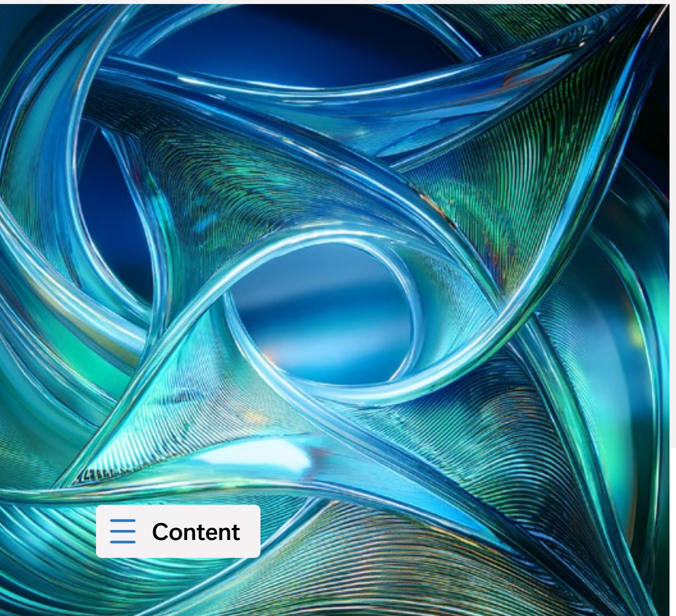
2024 Hackathons

738

hacks on AI included a focus on responsible AI

Map: Identifying risks

Mapping and prioritizing risks so that we can respond quickly and effectively is a critical component of our risk management approach. Tactics we use to identify and prioritize AI risks include threat modeling,[11] responsible AI impact assessments, customer feedback, incident response and learning programs, external research, and AI red teaming.[12] The mapping process informs decisions about planning, mitigations, and the appropriateness of an AI application for a given context. Equally important is our ability to remain flexible and responsive to new or previously unforeseen risks that arise at any stage of development or deployment, including post-deployment.



Advancements in red teaming operations, research, and tooling

Part of our process of mapping potential risks of AI systems and models includes red teaming, which is the process of simulating adversarial attacks to identify risks. For pre-deployment red teaming of our highest risk AI systems and models, we leverage the expertise of Microsoft's AI Red Team (AIRT), a centralized team of professional red teamers that operates independently of product teams. Guided by tools and resources developed by expert red teamers, product teams across Microsoft also perform pre-deployment red teaming of their AI systems and models.

In 2024, AIRT conducted 67 total operations across Copilots and models, including every flagship model added to the Azure OpenAI Service and every Phi model released. 2024 was also an unprecedented year for the introduction of new modalities. AIRT conducted red teaming on several new modalities, including image-to-image, video-to-text, text-to-video, and text-to-audio.

In a white paper titled "Lessons learned from red teaming 100 generative AI products,"[13] the team highlights more details on their operations, including case studies across modalities, an overview of the AI threat model ontology they developed to identify key components of attacks and vulnerabilities, and eight key lessons they have learned over the past few years.

In 2024, AIRT partnered with product teams and conducted multiple research operations to understand emerging AI capabilities and risks, including exploring the contours of agentic AI red teaming and developing strategies for future operations. The team also expanded the coverage of seed prompts used in red teaming operations to include a broader set of languages, emotional expressions, and singing audio sets for single turn jailbreaks. Other improvements included the integration of new third-party tools, such as InspectAI[14] and Vivaria,[15] for cybersecurity exercises.

While keeping up with the pace of red team operations across the company, AIRT also focused on sharing their learnings by up-skilling 1,500 internal and external cybersecurity and AI professionals on red teaming best practices.

The team also continued to improve the Python Risk Identification Toolkit (PyRIT), an open-source red teaming tool released in February 2024 that has since grown to over 2,500 stars on GitHub.[16] Improvements made to PyRIT since its release include creating a centralized way to share prompts and datasets across operations and improving scoring and reporting capabilities. In April 2025, Microsoft announced PyRIT's integration with Azure AI Foundry.[17] Customers using the AI Red Teaming Agent in Azure AI Foundry can simulate adversarial attack techniques and generate red teaming reports that help track risk mitigation improvements throughout the AI development lifecycle.

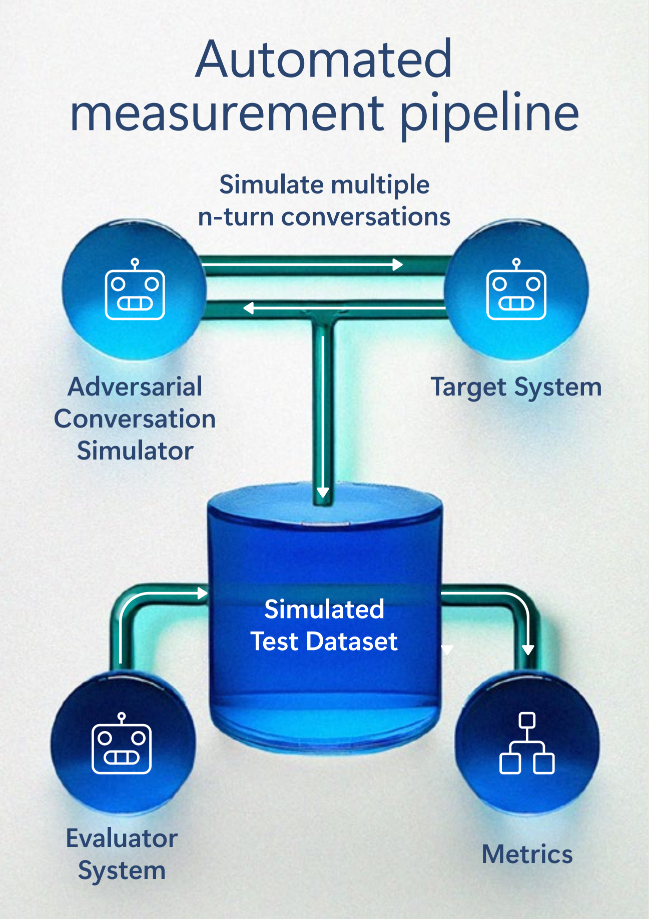
Measure: Assessing risks and mitigations

At Microsoft, the intentional collaborations we nurture between engineering, research, and policy are foundational to our practice of responsible AI. These collaborations are particularly important when it comes to advancing the science of AI measurement and evaluation. Our ability to develop effective and valid risk measurement capabilities that move at the speed of innovation became more important in 2024 as AI capabilities and the creative ways they are used continued to grow more complex.

AI risk measurement helps us to prioritize mitigations and assess their efficacy. In addition to regularly updating our measurement methods, we also share resources and tools that support the measurement of risks and risk mitigations with our customers.

Read more about our measurement tools on page 40

We continue to leverage the power of generative AI models to scale our measurement practices. Our automated measurement pipelines comprise three main components. The first component is the AI system or model that is being evaluated (target system). The second component is an AI model—usually an LLM, or in some cases, a multimodal model—that is instructed to interact with the target system by simulating adversarial user behavior (adversarial conversation simulator). The interactions between the target system and the adversarial conversation simulator generate outputs that make up a test dataset.


Automated measurement pipeline
Simulate multiple n-turn conversations
Adversarial Conversation Simulator
Target System
Simulated Test Dataset
Evaluator System
Metrics

The next step is to annotate the test dataset to identify which outputs contain content that is harmful or undesirable. For example, we may want to tag outputs that contain prompt injection attacks. To do this we use a third AI model (evaluator system) that annotates the test dataset based on policies and instructions developed by human experts. The accuracy of the AI annotations is compared against human annotations and improved as needed.

Finally, the annotated test datasets are used to calculate metrics that measure the proportion of the test dataset that contains harmful content. These metrics inform decisions about downstream mitigations that need to be applied.

In 2024, we leveraged expertise across research, engineering, and policy to make significant improvements to our measurement pipelines with the primary goal of expanding risk coverage across different modalities and risk types, enhancing the reliability of metrics generated, and leveraging new approaches to expose vulnerabilities.

For example, we expanded our measurement pipeline to cover the generation of critical election information and reproduction of protected materials. Our broader approach to mapping, measuring, and managing AI-related risks for 2024 elections is covered in more detail in the "Manage" section of this report.

Read more about our elections work on page 21

Our testing coverage for protected materials included content such as song lyrics, news, recipes, and code from public, licensed GitHub repositories.

We also expanded measurement support for image generation and understanding for sexual, violent, and self-harm content and content related to hate and unfairness. Furthermore, with increased support for audio modalities in the latest releases of generative AI models, we expanded measurement support for audio interactions by adding a transcription layer and running the text output through our measurement pipelines.

To improve the reliability of our metrics, we leveraged several prompt engineering techniques to optimize the performance of the annotation component of our measurement pipeline. To better measure vulnerabilities, we used adversarial fine-tuning to generate prompts that are more effective at revealing potential vulnerabilities in the system.

Looking ahead, we are integrating more advanced adversarial techniques and attack strategies to systematically measure vulnerabilities that could be exploited by malicious actors. We also plan to improve our evaluators for accuracy and support granular metrics, which in turn will empower our customers by improving their own interpretability and providing transparency through scorecards. We will continue to expand our testing risk coverage while refining our existing evaluations across various settings, newer models, modalities, and tools. We will also continue fostering collaborations with Microsoft Research to incorporate the latest advances in the science of AI risk evaluation into our tools and practices. This includes building measurement frameworks to better understand, interrogate, and compare measurements comprehensively through multiple lenses.

Research

Applying a social science lens to AI risk measurement

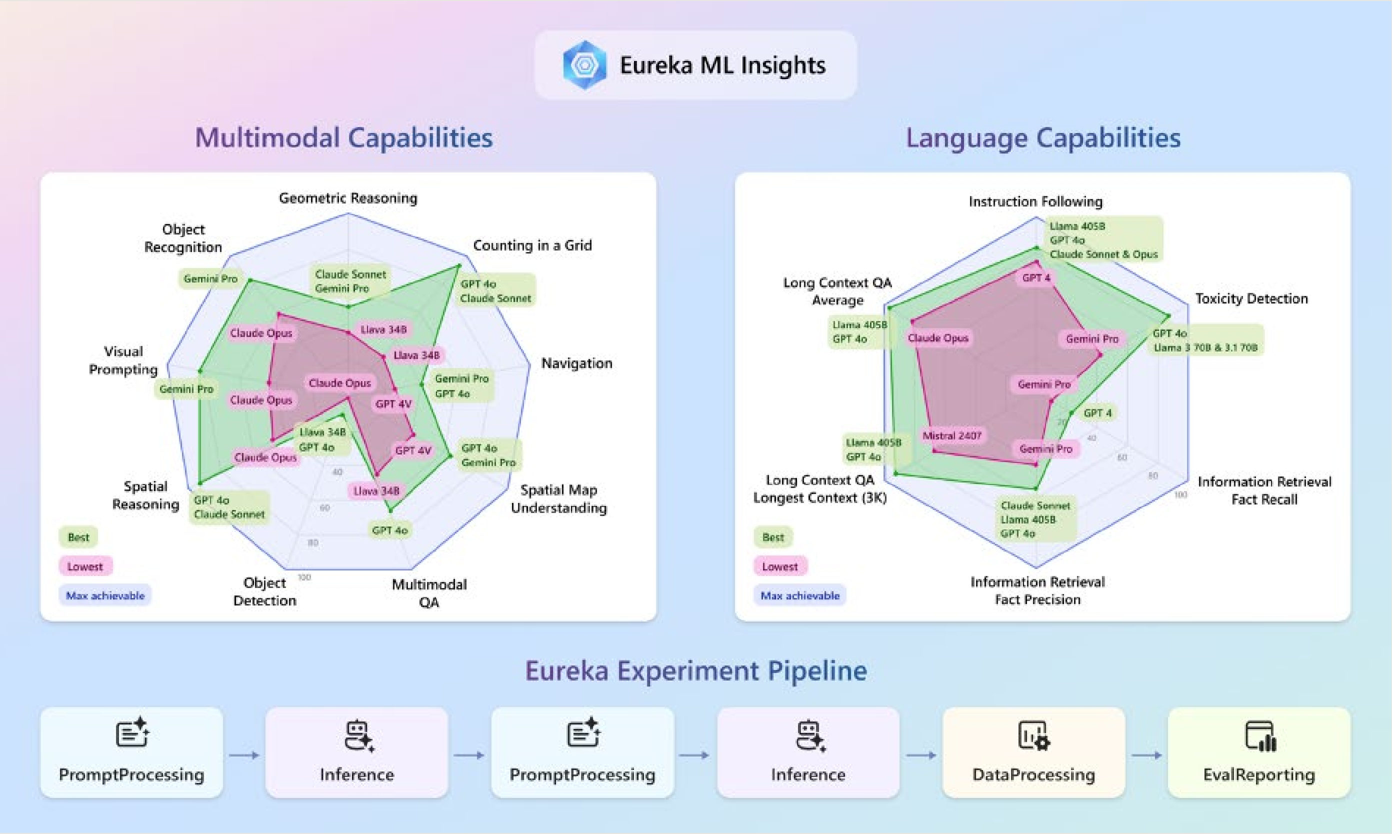
In 2024, the Sociotechnical Alignment Center (STAC)[18]—a team of researchers, applied scientists, and linguists within Microsoft Research—collaborated with researchers in the AI & Society Fellows Program to publish a paper that outlines a measurement framework grounded in practices that emerged in the social sciences.[19] The researchers argue that unlike more narrowly scoped measurement tasks (such as those involved in supervised machine learning systems), measurement tasks for generative AI systems often require measuring more complex, nuanced, and contested concepts.

STAC's proposed framework emphasizes the critical step of systematizing, or clearly defining, complex and nuanced concepts, amounts, populations, and instances before operationalizing measurements.[20] This provides clarity on what is being measured; enables stakeholders to better understand, interrogate, and compare measurements; and informs how effective risk mitigations are designed.

STAC's four-part measurement framework consists of the following components:

1. **Risk systematization:** Expert-led creation of a comprehensive description of the risk being measured, which informs downstream internal policy development.
2. **Datasets:** Systematic creation of test datasets that support different types of measurements and a scenario simulator for interacting with the system being evaluated.
3. **Annotation:** Manual and automated annotation of the outputs of the system being evaluated.
4. **Metrics:** Aggregation of the annotated outputs to create final measurements.

Additional research contributions from STAC in 2024 included methods to validate tests and guide decision-making during evaluation design. Validation tests ultimately ask the question: are we really measuring what we sought to measure? Each of the components in the framework needs to be validated using various lenses drawn from measurement theory in the social sciences. To help guide decision-making and provide a structure for comparing different evaluations, STAC has also proposed a set of general dimensions (evaluation setting, task type, input source, interaction style, duration, metric type, and scoring method) that capture critical choices involved in generative AI evaluation design.[21]



Addressing longstanding gaps in benchmark evaluations

In 2024, Microsoft Research's AI Frontiers lab worked on identifying and addressing a number of the prevailing challenges to current AI benchmark evaluation practices, including benchmark saturation and the lack of transparency in evaluation methods. To address these challenges and meet the need for more rigorous and nuanced evaluation of large foundation models, AI Frontiers developed Eureka, a reusable and open evaluation framework that aims to create transparency and reproducibility while standardizing evaluations of large foundation models.[22] AI Frontiers also released EUREKA-BENCH, a collection of benchmarks that state-of-the-art foundation models still find challenging to meet. These benchmarks represent fundamental but overlooked capabilities for completing tasks in both text and vision modalities.

The team used EUREKA-BENCH to conduct an analysis of 12 state-of-the-art models, providing in-depth insights for model comparison.[23] Their findings indicate that no single model currently excels across all capabilities, underscoring the importance of continued innovation and targeted improvements guided by detailed considerations of evaluations.

Manage: Mitigating AI risks

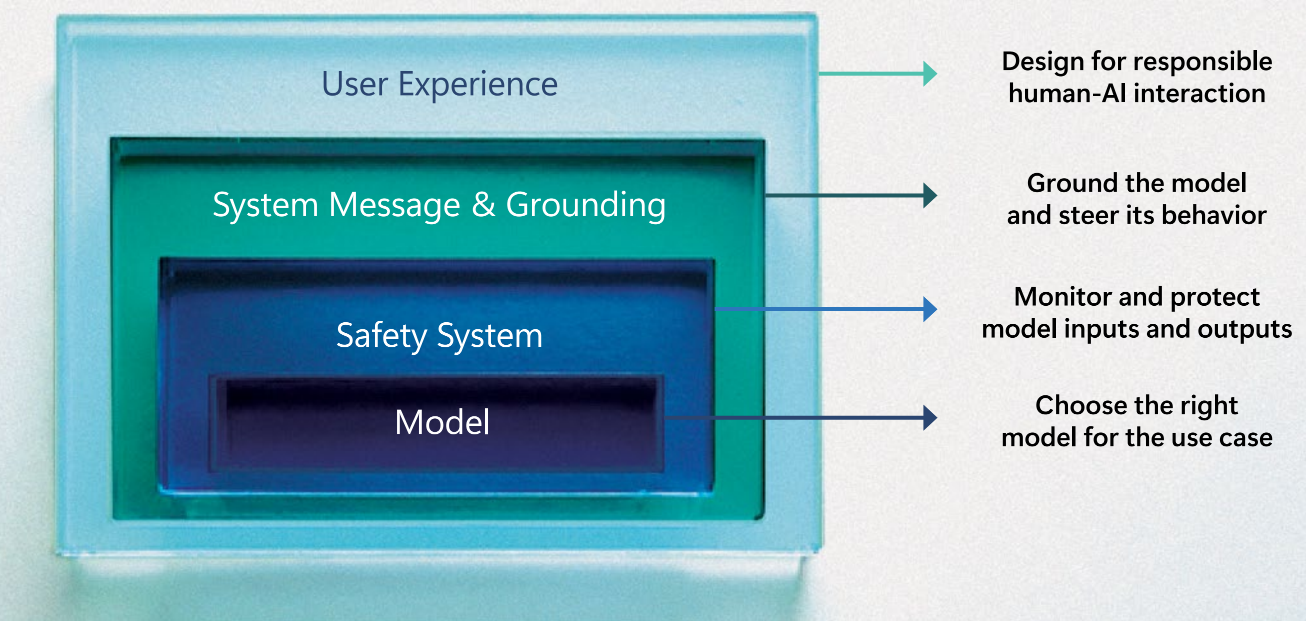
Once we've mapped and measured risks, we manage these risks across several layers of the AI technology stack to provide a "defense in depth" approach. At the platform level, safety measures such as content classifiers reduce risks by blocking potentially harmful user inputs and AI-generated content. At the application level, grounding a model's outputs with input data alongside safety system messages helps the application align with our Responsible AI Standard and user expectations. Even after we deploy AI systems, we continue to manage and mitigate risks through tooling and processes for ongoing monitoring, user feedback channels, incident response, and iterative improvements to the mitigation stack.

Strengthening Microsoft's AI safety stack

In 2024, we made significant performance improvements to the classifiers that detect and block sexual, violent, and self-harm content, as well as content related to hate and unfairness, for both user-generated and AI-generated text and images. We also made improvements to Prompt Shields, a unified API that detects and blocks jailbreak patterns in user inputs and indirect prompt injection attacks. After releasing Prompt Shields in March 2024, we have since expanded the API to detect gradient-based text attacks such as adversarial suffixes and attempts at agent hijacking via third party tools.[24]

We have also improved our classifiers that detect and block protected materials in AI-generated text such as song lyrics, news, and recipes. In September 2024, we extended these capabilities to flag AI-generated code that matches code from public, licensed GitHub repositories through Protected Materials for Code in preview in Azure AI Content Safety.

Read more on page 42 ›


User Experience
System Message & Grounding
Safety System
Model
Design for responsible human-AI interaction
Ground the model and steer its behavior
Monitor and protect model inputs and outputs
Choose the right model for the use case

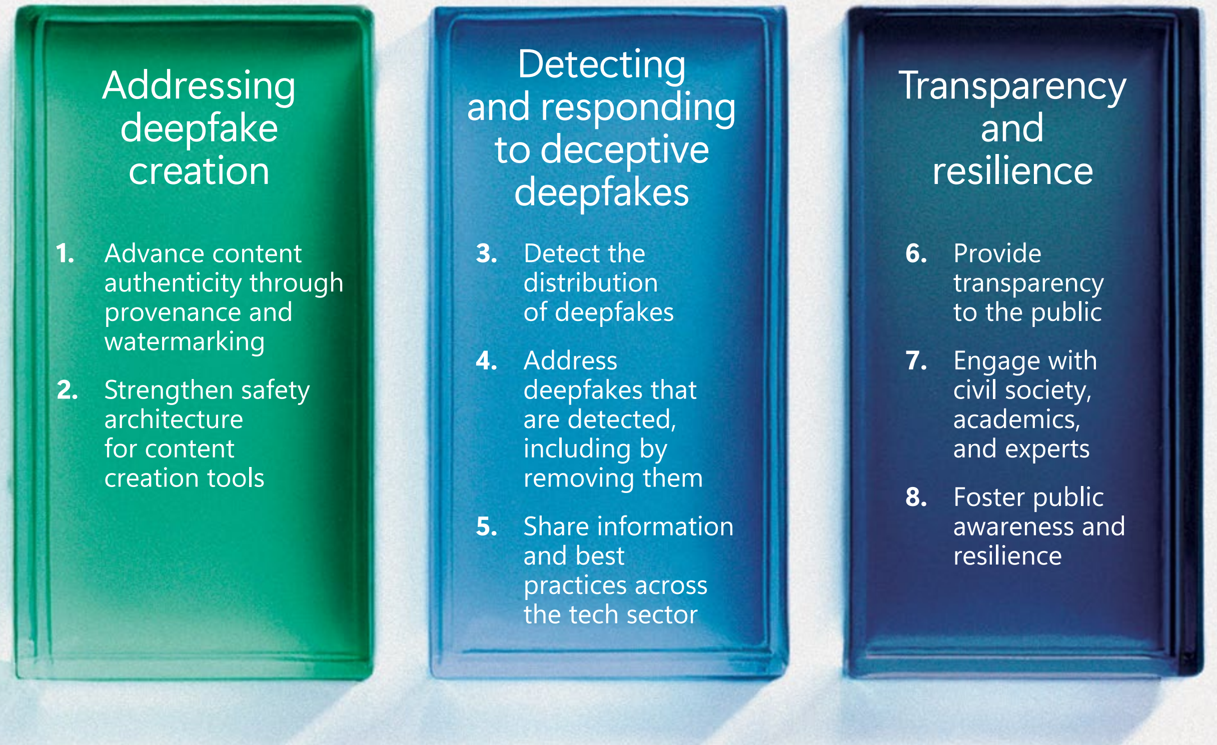
Managing AI-related risks in 2024 elections

In 2024, more people voted in elections across the world than at any other time in history. As a leading technology company, we know that we have a responsibility to take steps to prevent the creation and dissemination of deceptive AI-generated election content. Our proactive measures taken in partnership with governments, nonprofit organizations, and private sector companies globally helped to make sure that the world stayed ahead of AI-related threats to 2024 elections.

In February 2024, Microsoft joined twenty-six other leading technology companies in signing the Tech Accord to Combat Deceptive Use of AI in 2024 Elections, which are outlined below.[25]

Microsoft's efforts to live up to the commitments included in the Tech Accord started long before 2024. In 2021, we co-founded the Coalition for Content Provenance and Authenticity (C2PA) to develop an open technical standard for establishing the provenance—the source and history—of digital content, including non-AI-generated images, audio, and video. By 2023, leveraging the C2PA standard,

Tech Accord to Combat Deceptive Use of AI in 2024 Elections



Addressing deepfake creation

1. Advance content authenticity through provenance and watermarking
2. Strengthen safety architecture for content creation tools

Detecting and responding to deceptive deepfakes

3. Detect the distribution of deepfakes
4. Address deepfakes that are detected, including by removing them
5. Share information and best practices across the tech sector

Transparency and resilience

6. Provide transparency to the public
7. Engage with civil society, academics, and experts
8. Foster public awareness and resilience

we were automatically embedding cryptographically sealed provenance metadata, also known as Content Credentials, onto all content generated by the DALL-E series of models in Azure OpenAI Service, Bing Image Creator, Microsoft Copilot, Microsoft Designer, and Microsoft Paint.

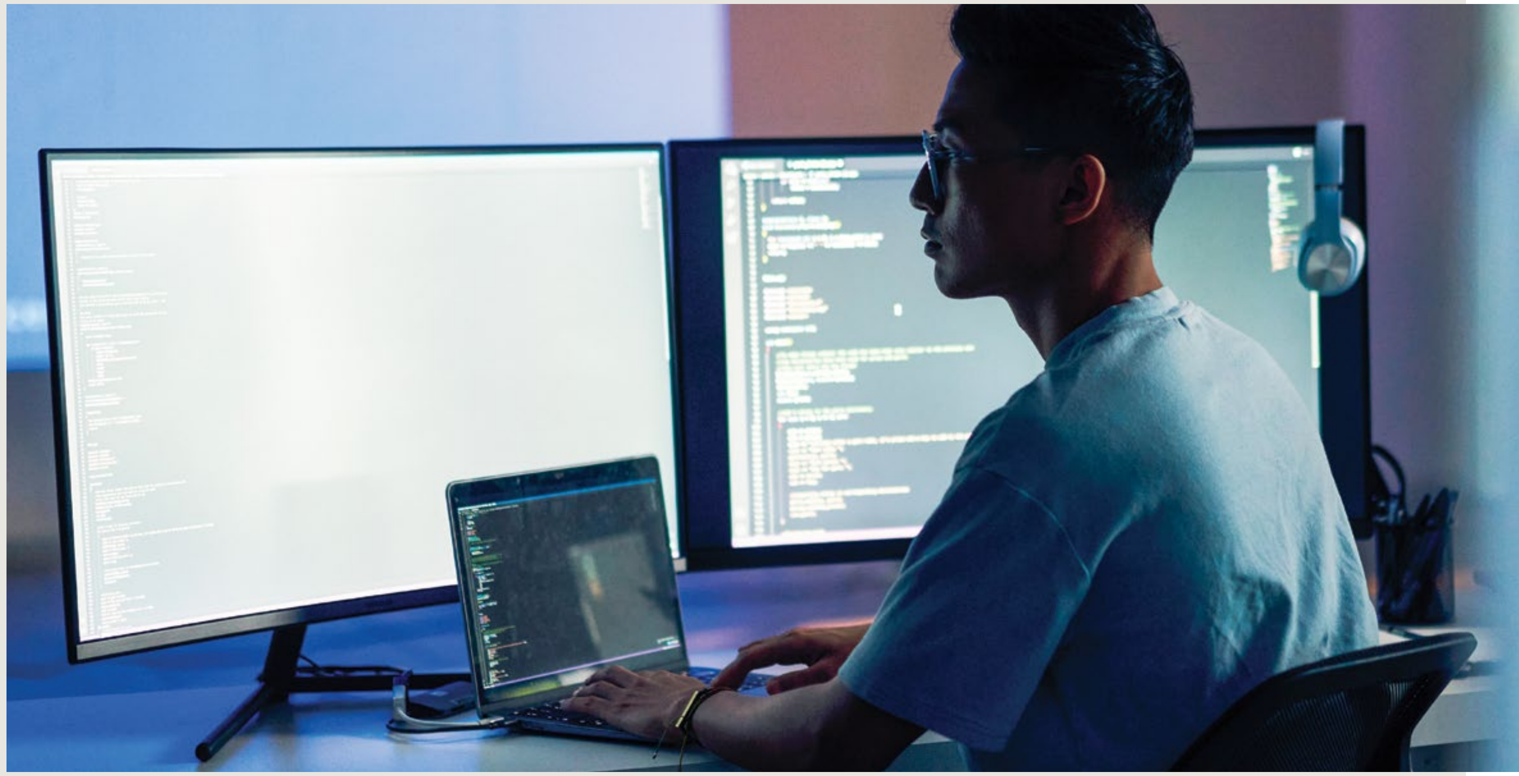
Throughout 2024, our efforts to manage AI-related risks in the 2024 elections centered around strengthening our AI safety architecture specifically for election-related risks and targeted efforts to detect and respond to reports of AI-generated deepfakes. On the safety architecture front, we focused efforts around managing two central risks:

1. The creation and dissemination of deceptive images of political candidates
2. Efforts to mislead voters about critical election information

To further these goals, we created specific product implementation guidance and tools to help teams across Microsoft mitigate the risk of their AI systems being used to generate deceptive images of candidates in key 2024 elections. If users sought critical election information—defined as factual aspects of political elections that could be clearly demonstrated to be true or false (e.g., where to vote, when to vote, and who was running)—Microsoft generative AI tools were instructed to provide a demonstrably reliable answer, directing users to specific authoritative sources (such as an election authority for that election); to redirect users to other tools such as Bing search that would surface authoritative sources; or refuse to answer the question if other approaches were not feasible. If users sought to generate an image depicting a political candidate of a 2024 election by name, Microsoft generative AI image tools were instructed to refuse to generate that image.

In February 2024, we created a site for candidates and election authorities to report election-related deceptive AI content like deepfakes appearing on Microsoft consumer services.[26] We continue to design and host trainings and briefings and send out pre-election communications to election offices, political parties, and campaigns globally to enable their use of the site. Beyond responding to reported events, we solicited the support of the Microsoft Threat Analysis Center (MTAC) whose mission is to detect, assess, and disrupt foreign cyber-enabled influence threats to Microsoft, its customers, and democracies worldwide. MTAC also partnered with the AI for Good team to develop technical capabilities to better detect deepfakes. In 2024, MTAC publicly published eight reports focused on nation-state actors and election interference.[27] The intelligence gathered by MTAC helped provide a broader view of the adversary threat landscape and enabled us to proactively combat deceptive AI-generated content.

While the 2024 elections are behind us, the threat that deceptive AI content could pose to elections around the world is far from over. At Microsoft, we recognize that we need to take a whole-of-society approach to address the risk of bad actors using AI and deepfakes to deceive the public. This is why we invest in open standards like C2PA and share the insights we gather through MTAC openly. It's why we also launched a fund with OpenAI to increase AI education among voters and vulnerable communities.[28] Microsoft will continue to develop our technology and policies, as well as work with other stakeholders globally, to ensure that we uphold the foundational principle of free expression for citizens in the United States and around the world.



Microsoft's approach to AI incident detection and response

The incident detection and response work that happens after the release of an AI product deserves just as much attention and planning as the work that happens leading up to the release. For the past two decades, the Secure Development Lifecycle (SDL) that Microsoft has implemented across the company has included a response phase that focuses on handling unforeseen issues and applying these learnings to future releases. We apply the infrastructure, processes, and best practices developed over the years through SDL to our AI systems.

Across Microsoft, product teams are required to put in place repeatable processes to collect user feedback and to triage and address issues that arise after the release of an AI system. Teams are also required to build feedback collection mechanisms within their products so users can more easily report concerns. When possible, teams employ automation to enable quick action on well-understood problems.

Potential incidents can be detected through various means, including automated detection, reports to the Microsoft Security Response Center's (MSRC) researcher portal, and employee reports,[29,30] which can also be anonymous. All such reports are triaged and assessed by expert teams. If these concerns are assessed to warrant an incident response, appropriate teams are assembled and coordinated by response specialists, who manage mitigation, root-cause analysis, and communication. After an incident, a postmortem analysis distills learnings from the event, which are both folded directly into making changes to improve system robustness and studied to identify patterns and themes which in turn shape our policies, practices, and priorities, including updating the bug bar used by the MSRC to triage concerns.

We also invest in crisis management to drive more consistency and efficiency in how we detect and respond to major AI issues and incidents. During major incidents, teams benefit from the expanded capacity of specialized roles like crisis managers, forensic investigators, and communications managers. As part of the product release cycle, we work with service teams to ensure they are connected to these central processes.

Looking across AI incidents

Security research and incidents provide valuable insights into how we can improve the engineering and operation of our AI services. In 2024, we formalized a process that brings those insights together to educate our engineering teams and update our engineering processes. We draw insights from every reported issue, each incident we experience, and the experiences of other AI labs in the threat landscape, including nation-state actor analyses reported by Microsoft's Threat Analysis Center.

Some key insights and learnings we gained through incident response in 2024 include the following:

- No incident-level events in 2024 were a result of AI system malfunctions or issues arising during benign use. Every incident included patterns of malicious use where actors were actively trying to bypass security measures or misuse Microsoft AI products or services.
- Threat actors identified in 2024 exhibited varying levels of sophistication, ranging from actors who worked in isolation to individuals who coordinated across a network of actors working towards the same goal.
- Threat actors often exploit differences in safety systems across different AI products and services, making it increasingly important to share the latest innovations in AI safety across the tech industry and with customers.
- Not only are threat actors circumventing safety systems built around generative AI systems and models, they are also using generative AI as a tool (e.g,. using AI to power spear phishing attacks at scale[31]), as they would PowerShell[32] or mimikatz.[33] Defenders are best equipped to think of AI as another tool in attackers' toolboxes.

We also draw insights from externally reported issues. When an issue is confirmed in an AI service, we open companion cases for other services using similar models to evaluate any impact. We also looks broadly across reported issues to synthesize themes, which are then used to educate product and service teams, update engineering processes to catch them earlier, and update policies as applicable.

Taking legal action against cybercriminals misusing AI

In January 2025, a complaint unsealed in the Eastern District of Virginia revealed that Microsoft's Digital Crimes Unit observed that an international threat–actor group had developed sophisticated software to exploit exposed customer credentials scraped from public websites.[34] In February, Microsoft filed an amended complaint that named four of the primary developers of these malicious tools.[35] These individuals, part of a global cybercrime network, unlawfully accessed accounts with certain generative AI services. They then altered the capabilities of these services and resold access to other malicious actors, providing detailed instructions on how to generate harmful and illicit content.

Upon our discovery of this activity, Microsoft revoked cybercriminal access, put in place countermeasures, and enhanced our safeguards to further block such malicious activity in the future. The court order also enabled us to seize a website instrumental to the criminal operation so that we could gather crucial evidence about the individuals behind these operations, decipher how these services were being monetized, and disrupt additional technical infrastructure we found.

Seizing this infrastructure allowed us to effectively disrupt a cybercriminal network and create a powerful deterrent impact among its members. We take the misuse of AI seriously and remain committed to protecting users by embedding robust AI guardrails and safeguarding our services from illegal and harmful content.

Controls to promote responsible use and for limited access services

Building on the responsible AI policies, processes, and practices that guide how we develop and deploy AI technologies, we have implemented other mechanisms to promote the responsible use of our technologies by our customers. For example, our enterprise contracts incorporate our AI Services Code of Conduct,[36] which requires our customers to implement responsible practices (such as human oversight and access controls) and prohibits using our AI services in ways that inflict harm on individuals, organizations, or society, or affects individuals in any way that is otherwise prohibited by law.

In addition to contractual obligations that apply to all of our AI services, we apply limited access restrictions for some of our AI services, such as our facial recognition technology, the custom neural voice AI-generated speech feature, and full configurability of content safety filters in Azure OpenAI Service.[37] Use of these services is only available to approved customers and partners who meet a combination of controls calibrated to the risks of each service.

These controls can include use case registration and pre-approval of use cases or restrictions on prohibited uses. For example, limited access restrictions require government agencies to meet strict criteria before gaining access to facial recognition services,[38] and prohibit the use of the custom neural voice feature to impersonate political figures in a way that could mislead the public. Customers may also be required to re-verify the information they submitted to gain access to these limited access services remains accurate, complete, and up-to-date.

Section 2

How we make decisions about releasing generative AI systems and models

Throughout 2024, we continued to refine our pre-deployment oversight processes, which include our deployment safety process for generative AI systems and models and Sensitive Uses and Emerging Technology program. In this section, we highlight progress we have made in improving these processes, lessons we have learned, and examples of AI systems and models that went through these review processes.

Deployment safety for generative AI systems and models

Before deploying their generative AI applications, teams review their risk management approach with experts across the Responsible AI community. These experts provide further recommendations and requirements grounded in our responsible AI policies, with an eye towards maintaining a consistent risk management approach for generative AI releases across the company.

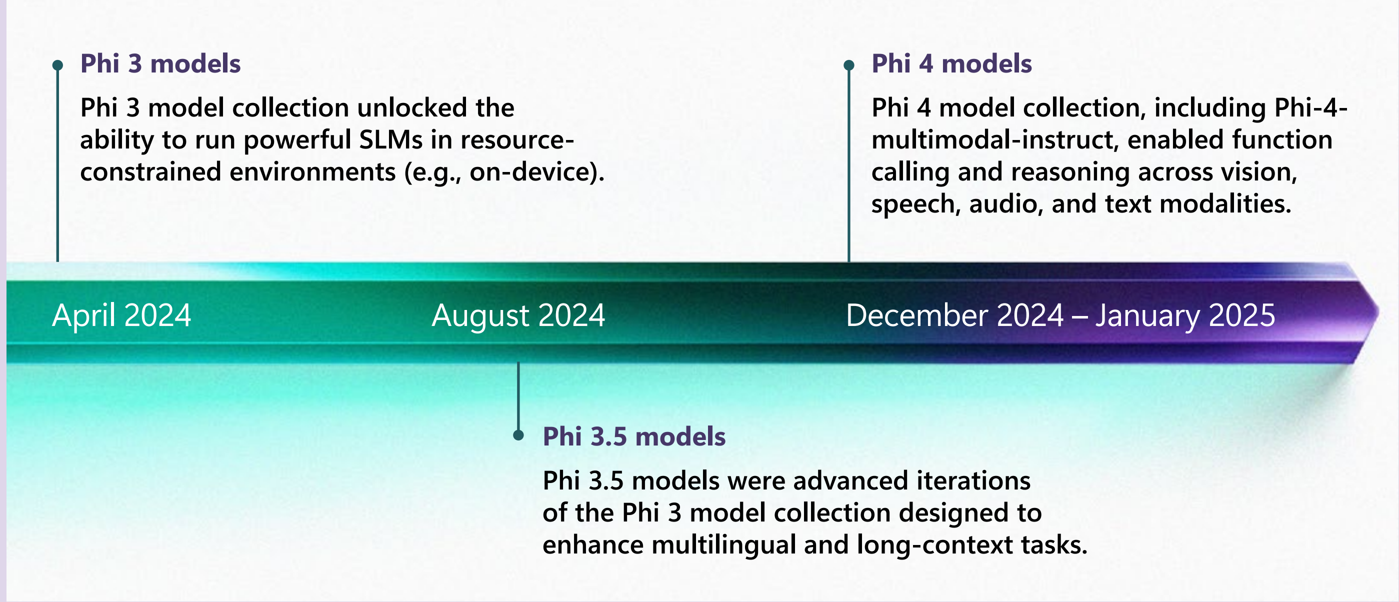
In 2024, there were more than 1,300 generative AI cases submitted to receive guidance from experts across the Responsible AI community. As more powerful, and in some cases, more cost-efficient generative AI models became available for product teams to use, we saw more cases of product teams upgrading the underlying models powering their generative AI applications, leveraging multimodal capabilities, and expanding support for additional languages.

To best manage the high volume of cases and maintain a consistently high bar for release, we developed resources that include tools, best practices, and instructions on how to map, measure, and manage risks. We also established a wider network of employees, including Responsible AI Champs, to help prepare product teams ahead of pre-deployment reviews with our experts. These preparation efforts often included helping teams take advantage of central tools and best practices and develop appropriate documentation.

Learnings gleaned across cases in 2024 helped to inform and prioritize gaps in policy, practices, and tools. In the Map, Measure, and Manage sections of this report, we discuss in greater detail the progress and changes we made to our centralized responsible AI practices and tools to help product teams safely deploy their generative AI applications and models.

One of the ways we updated our pre-deployment oversight process for generative AI was by launching an internal workflow tool to further support responsible AI documentation and review processes. This tool is designed to centralize the various responsible AI requirements outlined in the Responsible AI Standard into one central workflow to reduce unnecessary toil and make it easy for our teams to complete their responsible AI work throughout the development lifecycle. In mid-September, we released a reporting dashboard for our internal teams that allows us to track cross-company onboarding to the tool and completion of requirements. Regular updates are also shared with the Responsible AI Governance Community to support their use of the tool and keep them informed of the ongoing improvements we are making to the tool.

Case study: Safely releasing the Phi family of small language models



Phi 3 models

Phi 3 model collection unlocked the ability to run powerful SLMs in resource-constrained environments (e.g., on-device).

Phi 4 models

Phi 4 model collection, including Phi-4-multimodal-instruct, enabled function calling and reasoning across vision, speech, audio, and text modalities.

April 2024

August 2024

December 2024 – January 2025

Phi 3.5 models

Phi 3.5 models were advanced iterations of the Phi 3 model collection designed to enhance multilingual and long-context tasks.

Microsoft's Phi family of small language models (SLM) are designed to deliver high-quality outputs while maintaining a lightweight architecture for cost-effective deployment.[39]

As the Microsoft team developing the Phi models has continued to push the boundaries of SLM capabilities, they have refined their approach to implementing responsible AI in the Phi models. By 2024, the Phi model team had developed a "break-fix" framework grounded in Microsoft's broader AI risk management framework. The "break-fix" framework includes the following five iterative steps to build safer AI models:[40]

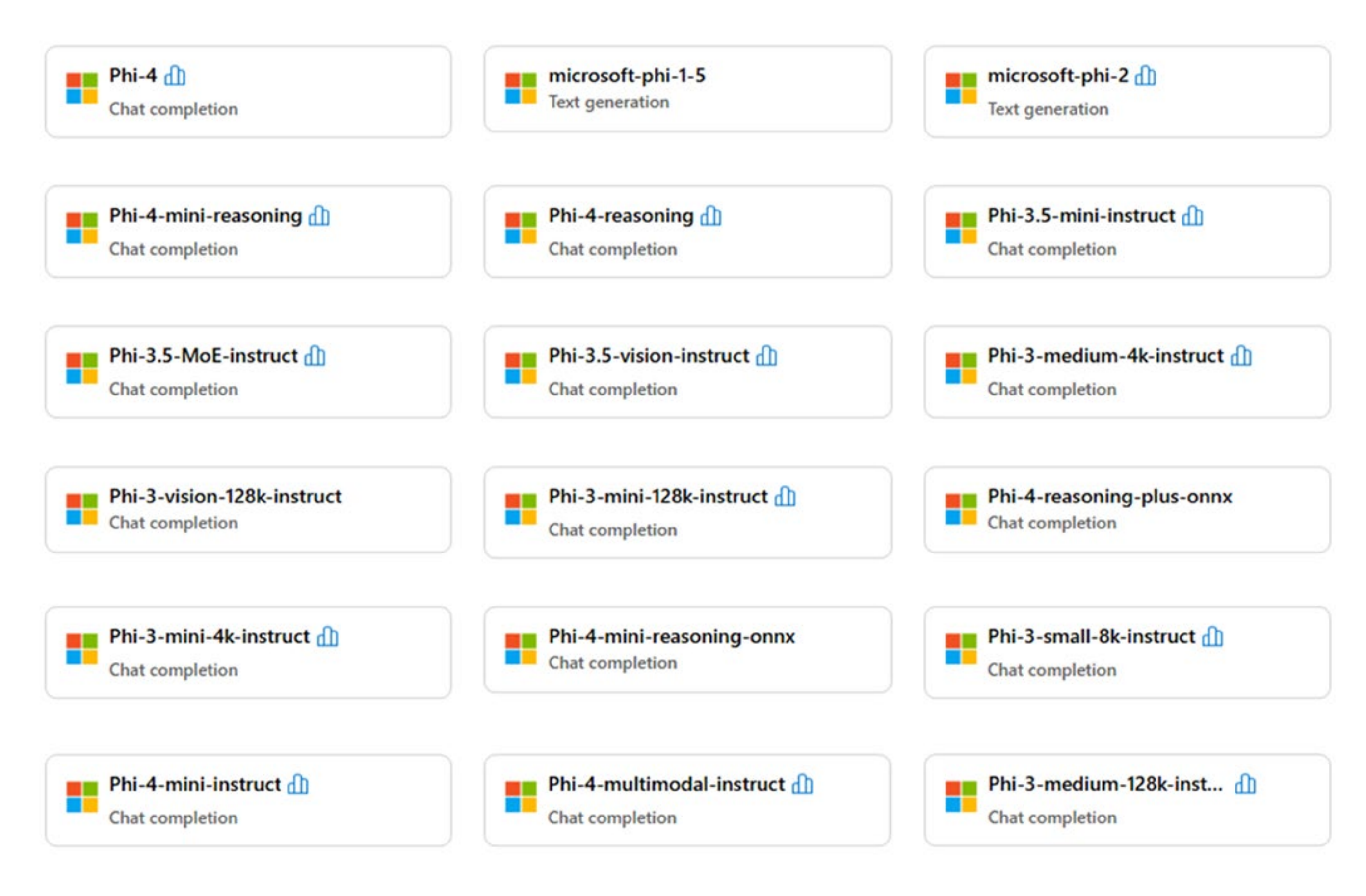
1. **Dataset curation:** Curate existing publicly available datasets with various modifications and synthetically generate additional datasets based on vulnerability insights.
2. **Post-training:** Leverage the safety datasets along with standard preference datasets in both the supervised fine-tuning and direct preference optimization stages of post-training.
3. **Quantitative and qualitative evaluations:** Perform a wide range of responsible AI evaluations to select release candidates for further red teaming.
4. **AI red teaming:** Perform red teaming on model release candidates.
5. **Vulnerability identification:** Based on the evaluations and red team findings, identify potential vulnerabilities to inform further safety dataset curation and safety post-training.

We describe below how the "break-fix" framework applies across the map, measure, and manage steps, highlighting key risk areas for each of the aforementioned Phi releases that were identified and mitigated by the Phi model team.

Map

In partnership with the AI Red Team, the Phi model team mapped key risks associated with each model release in an iterative fashion. As the Phi models are designed to be general purpose and could be used in a wide range of contexts, adversarial probing of the models focused on novel capabilities and risk areas that could not otherwise be covered by existing automated measurement pipelines or publicly available benchmark datasets.

- **Phi 3:** Red teaming for Phi 3 models covered a wide range of risk areas for both the vision and text models. Adversarial probing efforts included content related to current events, phishing and cybersecurity, hate speech, sexual content, violence, and more. For Phi 3 Vision, red teaming also focused on unique risks associated with the model's ability to process images. These efforts included probing the model's susceptibility to jailbreak techniques and its ability to identify faces of individuals, infer sensitive attributes from images of people, or be used to read captchas.
- **Phi 3.5:** The Phi 3.5 collection of models added multilingual support to the repertoire of Phi capabilities. The focus on language support led to significant advances in multilingual and culturally informed red teaming practices, such as the use of low-resource languages and script systems to bypass safeguards. These practices not only were key to the hardening of Phi 3.5, but became core to our practices for testing other models and systems, especially with the emergence of audio-to-audio models later in the year.
- **Phi 4:** Red teaming for the Phi-4 models focused on the newest audio and speech input modality across eight languages and multiple risk areas. The models were found to be more susceptible to providing undesirable outputs when attacked with context manipulation or persuasive techniques. These findings applied to all languages, with the persuasive techniques mostly affecting French and Italian. The models also showed the ability to infer sensitive attributes (e.g., personality characteristics and country/region of origin) from speech and audio inputs, a key risk that we subsequently mitigated.



Phi-4
Chat completion

microsoft-phi-1-5
Text generation

microsoft-phi-2
Text generation

Phi-4-mini-reasoning
Chat completion

Phi-4-reasoning
Chat completion

Phi-3.5-mini-instruct
Chat completion

Phi-3.5-MoE-instruct
Chat completion

Phi-3.5-vision-instruct
Chat completion

Phi-3-medium-4k-instruct
Chat completion

Phi-3-vision-128k-instruct
Chat completion

Phi-3-mini-128k-instruct
Chat completion

Phi-4-reasoning-plus-onnx
Chat completion

Phi-3-mini-4k-instruct
Chat completion

Phi-4-mini-reasoning-onnx
Chat completion

Phi-3-small-8k-instruct
Chat completion

Phi-4-mini-instruct
Chat completion

Phi-4-multimodal-instruct
Chat completion

Phi-3-medium-128k-inst...
Chat completion

Measure and manage

For each of the Phi model releases in 2024 and early 2025, the team used insights from red teaming to identify a combination of both internally developed measurement datasets and publicly sourced benchmark datasets to assess the prevalence of risks and the efficacy of mitigations applied. Following the break-fix framework, the team iterated through several rounds of measurement and mitigation steps, working closely with responsible AI experts to determine when the models were sufficiently safe for release. They compared the release candidates to other comparable models already on the market and carefully considered deployment methods (in this case, the model was both open source and made available through Azure AI Foundry).[41]

Below we lay out steps the team took to measure and manage novel risks associated with each of the Phi model releases in 2024 and early 2025.

- **Phi 3:** For the Phi 3 models, the team leveraged a combination of internal and publicly available safety benchmark datasets to assess release candidate models through every break-fix round. These included simulated adversarial conversation sets covering risk areas such as groundedness and susceptibility to jailbreak techniques, and external benchmark datasets such as XSTest that was used to measure both appropriate and inappropriate refusal rates of the models.[42] Red teaming and benchmark evaluation efforts informed subsequent safety data curation and safety post-training as part of the break-fix approach. After several break-fix iterations, the team observed an average of 75% reduction in the amount of harmful content generated by Phi 3 models, including Phi 3 Vision, which had significant improvements for risks such as reading captchas and susceptibility to jailbreak techniques.[43]

- **Phi 3.5:** The availability and reliability of multilingual safety evaluation and performance benchmark datasets continues to be an industry-wide challenge. To compensate for the limitations of any one approach or dataset, the team used a combination of measurement approaches, including simulated adversarial conversation sets in multiple languages and publicly available multilingual benchmark datasets. These datasets were used to assess Phi-3.5 models' susceptibility to jailbreaking and their propensity to produce undesirable outputs across multiple languages and risk categories.

  For this release, findings across the various evaluation methods indicated that safety post-training efforts had a positive impact across multiple languages and risk categories, as observed by higher rates of the model refusing to output undesirable outputs, including when jailbreak techniques were leveraged. However, red teaming indicated the model sometimes generated refusals in English, even when the request for undesirable output was in another language. These findings highlighted the need for industry-wide investment in the development of high-quality safety evaluation datasets across multiple languages that include low-resource languages and risk areas that account for cultural nuances where those languages are spoken.

- **Phi 4:** When building the Phi 4 models, the team employed several new techniques to unlock unique capabilities and address risks observed in prior releases. They used additional data to enable audio and speech modalities, designed a new architecture for efficiency, enabled a larger vocabulary for multilingual support, expanded safety post-training data to include more languages, and applied new post-training techniques to enhance the models' ability to follow instructions and use external tools such as APIs. For this release, as for prior releases, the team ran both simulated adversarial conversation sets and publicly available benchmark datasets, often repeating similar evaluation approaches to prior releases for the same modalities. They found that for text and vision scenarios, Phi-4 models made significant safety gains when compared to prior releases, including robustness to jailbreak techniques and higher refusal rates to harmful prompts.

  For the audio modality available in Phi-4 Multimodal, in addition to simulated adversarial conversation sets, the team also evaluated performance differences across various demographic groups for speech-to-text transcription use cases. Their evaluations focused on gender and age groups across multiple languages and found very minimal differences. The team also used red teaming to test the model's ability to infer sensitive attributes from users' voices. They found that while the model inferred sensitive attributes in 27% of test prompts, the most common attributes being personality characteristics and country/region of origin, the use of a system safety message instructing the model not to infer sensitive attributes dropped this rate down to 0.4%.

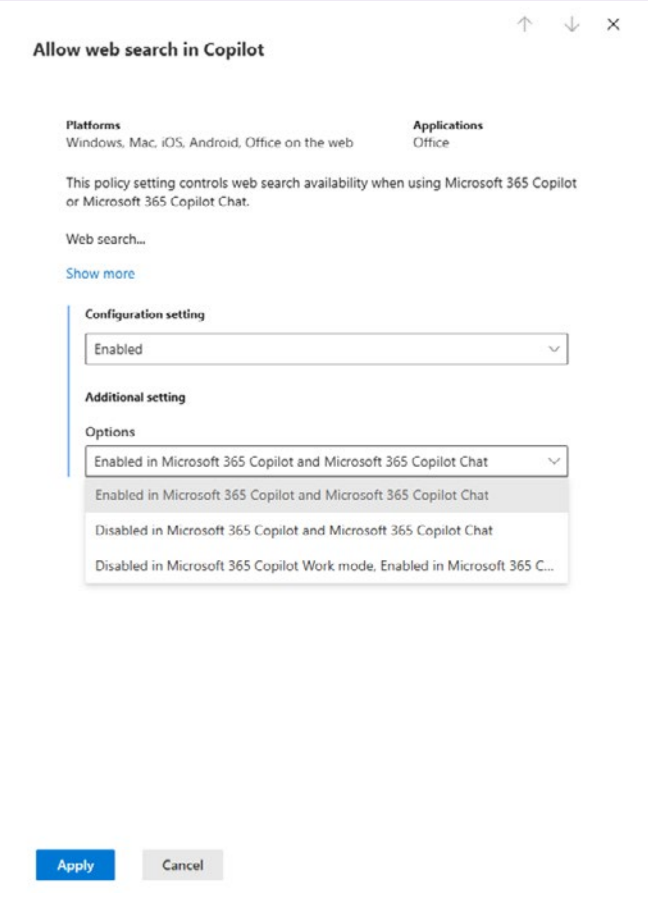
Additional details on testing results can be found in the technical reports and model cards that accompanied the Phi-3, Phi 3.5,[44,45,46] and Phi-4[47,48,49] model collection releases. These documentation practices play a key role in managing risks by empowering developers with the information they need to innovate responsibly with Phi models.

Case study: Improving web search controls for Microsoft 365 Copilot

Microsoft 365 (M365) Copilot is an AI-powered productivity tool that is integrated with M365 apps and services such as Word, Excel, PowerPoint, Outlook, Teams, and more.[50] It offers intelligent assistance tailored to each application. For example, in Word, M365 Copilot can help create documents; in Excel, it can suggest formulas; in Outlook, it can summarize email threads; and in Teams, it can summarize meetings. These integrations allow users to boost their creativity, productivity, and skills across different applications.

When a user enter a prompt, M365 Copilot uses a combination of models provided by the Azure OpenAI Service to generate responses in real time. To personalize responses, M365 Copilot utilizes content from Microsoft Graph, which includes a user's work emails, chats, and documents. It is core to the design of M365 Copilot that the application only displays data that the user has existing permissions to access. Responses may also be grounded in Internet-based content, which requires performing a web search based on the user's prompt.

As captured in the Transparency Note for M365 Copilot, the relevant product teams mapped, measured, and managed risks related to the use of the Copilot feature across M365 apps and services ahead of initial deployment and each subsequent release cycle.[51] They completed all required privacy and security reviews, which are core tenets of Microsoft's approach to building AI responsibly. They reviewed their AI risk management approach with responsible AI experts from across Microsoft ahead of each release cycle and established mechanisms to monitor customer feedback. Once M365 Copilot was deployed, one recurrent request from customers was greater control over web search capabilities in their interactions with M365 Copilot.



Allow web search in Copilot

Platforms
Windows, Mac, iOS, Android, Office on the web

Applications
Office

This policy setting controls web search availability when using Microsoft 365 Copilot or Microsoft 365 Copilot Chat.

Web search...

Show more

Configuration setting

Enabled

Additional setting

Options

Enabled in Microsoft 365 Copilot and Microsoft 365 Copilot Chat

Enabled in Microsoft 365 Copilot and Microsoft 365 Copilot Chat
Disabled in Microsoft 365 Copilot and Microsoft 365 Copilot Chat
Disabled in Microsoft 365 Copilot Work mode, Enabled in Microsoft 365 C...

Previously, access to web search in M365 Copilot was based on the M365 "optional connected experiences" setting enabled at the tenant level. This setting enables a single configuration that impacts a group of features, which led to the challenge that web search across M365 Copilot could not be configured separately.[52]

Through the feedback channels that the product teams established ahead of M365 Copilot's release, customers expressed the need for more granularity when making data privacy control choices and requested the option to turn web search on or off independent from other features.

Based on this feedback, in November 2024, Microsoft shipped the "allow web search in Copilot" setting, which allows IT administrators the ability to enable or disable web search in M365 Copilot for all users and user groups.[53]

If an IT administrator enables web search for the tenant, end users have the option to turn off web search by using the "web content" toggle across M365 Copilot available in each M365 app and service.

This more granular control over web search in M365 Copilot was well received by many customers, including customers across financial services, automotive, and health and life sciences industries.

In early 2025, to continue earning customers' trust, Microsoft successfully attained the ISO/IEC 42001:2023 certification for M365 Copilot and M365 Copilot Chat.[54] ISO/IEC 42001:2023 is the first international standard for AI Management Systems. It provides organizations with a structured framework to responsibly develop, deploy, and manage AI systems.

Attaining this certification confirms that an independent third party has validated Microsoft's application of the necessary framework and capabilities to effectively manage risks and opportunities associated with the continuous development, deployment, and operation of M365 Copilot and M365 Copilot Chat.

Microsoft was a key contributor to the conception and development of the ISO/IEC 42001:2023 standard. We remain committed to advancing internationally recognized standards that help to establish consistent practices, enhance accountability, and foster trust in AI technologies.

Sensitive Uses and Emerging Technologies program

Our Sensitive Uses and Emerging Technologies program continues to serve as Microsoft's review and oversight process for high-impact and higher-risk uses of AI. Through this process, we provide guidance for all types of AI systems, generative or otherwise.

AI systems and models developed or deployed by Microsoft, whose foreseeable use or misuse meets one of the following criteria are reported for Sensitive Uses review:

- Systems that could have a consequential impact on an individual's legal status or life opportunities, e.g., systems that prioritize access to healthcare services.
- Systems that could present the risk of significant physical or psychological injury, e.g., workplace safety alert systems in industrial environments.
- Systems that could restrict, infringe upon, or undermine the ability to realize an individual's human rights, e.g., computer vision-based systems deployed in public spaces that could restrict rights to assembly.

The Sensitive Uses review process culminates in requirements that often go beyond the baseline, generalized requirements outlined in the Responsible AI Standard and related policies. The Sensitive Uses and Emerging Technologies team consists of a multidisciplinary set of experts who provide hands-on counseling for high-impact and higher-risk uses of AI, as well as research and guidance for emerging issues or novel AI technologies. The team includes members with backgrounds in engineering, cybersecurity, product management, public policy, international relations, user experience research, data science, social sciences, and law. The team's expertise is further augmented by professionals from across our research, policy, and engineering organizations with expertise in human rights, social science, privacy, and security, who lend their expertise on complex sociotechnical issues as part of the Sensitive Uses Panel.

Throughout 2024, teams across Microsoft sought responsible AI guidance from the Sensitive Uses and Emerging Technologies team. The predominance of generative AI in these consultations—accounting for 77% of the cases—underscores its growing impact and the importance of addressing its challenges responsibly. The case studies that follow provide further insight into how the process works in practice.

77% of consultations focused on generative AI

Case study: Voice capabilities in Microsoft Copilot

Microsoft Copilot (Copilot) is an AI-powered chat experience that aims to boost both creativity and productivity. Users can interact with Copilot to, among many things, brainstorm ideas, conduct research, create images, or simply have an engaging conversation. Using advanced AI models, Copilot reasons across multiple modalities to go beyond answering basic queries and focuses on providing users with a more personalized AI experience. As part of the broader Copilot relaunch in October 2024, the team building the system sought to leverage the audio capabilities of GPT-4o to provide a more natural conversational experience known as "Copilot Voice."

Deploying an AI system with new modalities such as voice and audio introduces novel risks that warrant additional review and oversight. The Copilot Voice product team went through Sensitive Uses review, where they received hands-on consultation from the Sensitive Uses team. Through the review process, the team identified, measured, and mitigated risks associated with audio generation, as well as user voice inputs, before deploying Copilot Voice.

Mapping risks

The Sensitive Uses and Emerging Technologies team coordinated extensive red teaming conducted by both internal and external red teams. The red teams probed both the underlying GPT-4o model powering the Copilot Voice experience and the user-facing application with and without additional safeguards applied. The intention of these focused exercises was to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations related to voice and audio scenarios. This process helped the product team better understand how the system could be used by a variety of users and helped improve mitigations.

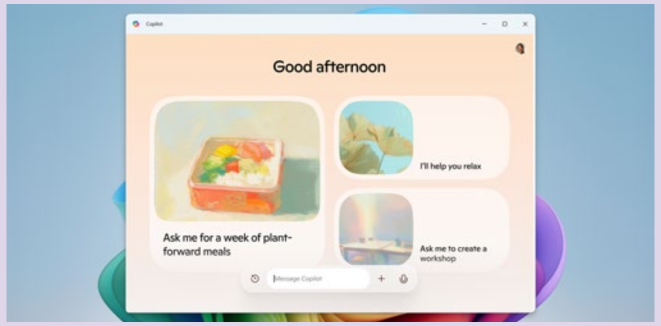
Measuring and managing risks

Taking signals from red teaming, the product team conducted further evaluations through partially automated measurement pipelines. The Sensitive Uses and Emerging Technologies team worked closely with the product team to design broad coverage of risk areas with a focus on voice and audio-enabled risks. This included risks related to the system's ability to reproduce or mimic a person's voice, susceptibility to voice-based jailbreaking techniques, generate third-party content, and more.

The measurement pipelines included a set of conversations, or interactions, with Copilot Voice collected from human evaluators and synthetic conversations generated with LLMs prompted to test policies in an adversarial fashion. Each of the newly developed conversation sets were annotated by human labelers who read the text content or listened to the audio output to validate the LLM-based evaluations.

These measurement efforts informed the development of a range of mitigations, including post-training, system prompts, and both input and output classifiers. For example, to mitigate the risk of the system reproducing or mimicking a person's voice, the product team implemented a mechanism to assess and block voice outputs from the system that diverge significantly from the set of voices the user can choose from.

Finally, the Sensitive Uses and Emerging Technologies team helped develop and review transparency documentation designed to equip users with information they need to use the product responsibly.[55] The transparency documentation includes information about the AI-powered feature's capabilities, limitations, intended uses, and best practices for use.


Copilot
Good afternoon
Ask me for a week of plant-forward meals
I'll help you relax
Ask me to create a workshop
Message Copilot

By design, cases submitted to the Sensitive Uses and Emerging Technologies team tend to be applications of AI with complex risk profiles. This means that cases are often fact-intensive and require bespoke engagement and guidance. However, we observed several common trends in Sensitive Use cases in 2024:

1. **Use of generative AI has continued to accelerate, with emphasis on multimodal and agentic applications.** 77% of Sensitive Use cases in the past year involved leveraging a generative model, deployments which increasingly have image- and audio-based capabilities. More recently, many of these cases have also involved agentic applications of AI, such as generative orchestration,[56] which involves coordination of information flows between user input across one or more models and across additional tools and sources of data.

2. **Health and life science projects make up an increasing share of cases submitted.** While the Sensitive Uses and Emerging Technologies team reviews cases across almost every industry, we receive a higher volume of health and life science submissions compared to any other domain. These projects typically involve administrative and decision-support tools for healthcare professionals and educational tools for patients. Deployment scenarios often meet two of our Sensitive Uses reporting criteria: Systems that could have a consequential impact on an individual's legal status or life opportunities (e.g., access to healthcare services), and systems that could present the risk of significant physical or psychological injury (e.g., systems designed to detect health conditions). Reviews for these features often focus on how the features can be evaluated for accuracy, both in the lab and once deployed, and how we can develop effective strategies to reduce risks from users over-relying on their outputs.

3. **The number of cases involving AI for scientific research has increased.** We received more cases involving AI for scientific research in 2024 than in prior years. Microsoft's research and development efforts in the sciences are driven by experts in quantum physics, computational chemistry, molecular biology, software engineering, and other disciplines. Their work increasingly makes use of AI innovations. For example, in 2024, Microsoft's AI for Science research organization[57] developed MatterGen,[58] a generative AI tool designed to accelerate materials discovery with potential application in broad domains including batteries, magnets, and fuel cells. Given the many potential downstream applications of these research advances, the Sensitive Uses reviews typically focus on developing a detailed threat model, which is used to build customized research safeguards and deployment mitigations.

In addition to the case consultation function, our Sensitive Uses and Emerging Technologies team develops early guidance for emerging AI technologies and risks. By identifying signals from Sensitive Uses cases, partnering with researchers and product strategists across Microsoft, and scanning the horizon of AI innovations, the team creates early-steer guidance ahead of broader, more formal policy development. This provides engineering teams with actionable recommendations when building with novel and emerging AI technologies. In 2024, this included drafting guidance for audio-based systems, image editing, and agentic AI systems. Early steer guidance also serves as a foundational resource for developing more formal internal policies and helps inform our public policy efforts around new AI models and applications.

Case study: Smart Impression

Smart Impression is an AI-powered feature within a suite of productivity tools for radiologists called PowerScribe One. In radiology reports, the "impression" of the report refers to the section that summarizes all clinically significant findings and recommendations to inform downstream patient care. Smart Impression uses a language model to compose draft impressions based on a radiologist's findings. By enhancing the speed and accuracy of composing these impressions, Smart Impression can improve efficiency and reduce the risk of burnout among medical professionals.

The use of generative AI in a healthcare decision support tool is a Sensitive Use case that warrants additional review and oversight. Through the Sensitive Uses review process, the product team received hands-on consultation to identify several key risks, including the possibility of generating impressions that don't align with findings included in the report—also known as ungrounded content.

To measure and evaluate the risk of generating ungrounded content, the product team conducted both automated and human evaluations for each product release cycle, relying on practicing radiologists for their expertise. Their measurement approach tracked various types of discrepancies, such as the tool mentioning a finding that was not in the reference impression or report, and omitting the anatomic location or position of a finding present in the reference impression, among other variations.

The product team put in place several layers of mitigation to align with the requirements issued by the Sensitive Uses and Emerging Technologies team, including:

- **User-initiated:** The Smart Impression feature is off by default, meaning it cannot generate content without being invoked by the user and, if invoked, its final output can be discarded or edited. The radiologist has the option to write their own impression or invoke Smart Impression to generate draft impressions via a button push or voice command.
- **Human review:** If the Smart Impression feature is invoked by the user to generate draft impressions, a message is displayed to indicate that there is AI-generated content that needs to be reviewed for accuracy, which also triggers a workflow that prompts the user to review and approve it. The draft impression cannot be submitted into a patient's medical record without review and signoff by the radiologist.
- **Feedback mechanism:** Radiologists can submit concerns or feedback about their experience with the tool through the user feedback feature, which is used to make ongoing improvements to Smart Impression.
- **Transparency:** The team developed transparency documentation to equip radiologists with the information they need to use the Smart Impression feature responsibly. This documentation includes information about the AI-powered feature's capabilities, limitations, intended uses, and best practices for use. The transparency documentation is made available to Smart Impression customers by default. Prospective customers can get access to the transparency documentation upon request.

In line with Microsoft's staged release approach, Smart Impression was released for private preview to a small set of practicing radiologists to gather real-world insights ahead of broader release efforts. The private preview included 40,000 radiology reports generated by over 100 radiologists from community care hospitals, private practices, and university hospitals in the U.S. In two rounds of focused interviews as well as three site visits, radiologists relayed that they found the results to be acceptable for use in real radiology interpretation scenarios and found the draft impressions to be helpful both in terms of time savings as well as reducing cognitive load.

Before public preview, the product team used insights from interviews, feedback submitted by radiologists through the user feedback feature, and analysis of generated impressions that were accepted as-is or with minimal edits to make changes and improve performance of the model. By the end of public preview, half of the AI-generated impressions were accepted as-is, and almost three quarters were accepted with only minor edits by practicing radiologists. Smart Impression is now available to radiologists across the U.S., and the team continues to monitor and address issues to improve outcomes for patients and healthcare professionals.

Section 3

How we support our customers in building AI responsibly

As developers and deployers of AI technology, we view it as our responsibility to support our customers in their own responsible AI journeys by equipping them with many of the tools and practices we use internally. We eagerly engage in dialogue with our customers to learn what's working and what's not, and how we can better support them in innovating responsibly.

AI Customer Commitments

Microsoft's AI Customer Commitments were first launched in 2023,[59] starting with our Customer Copyright Commitment. Our Customer Copyright Commitment outlines Microsoft's agreement to defend customers against claims of copyright infringement as a result of the output content of Copilot or Azure OpenAI Service, assuming required mitigations are put in place. In 2024, after listening closely to questions and feedback from our partners, we extended our Customer Copyright Commitments to include our reseller partners. This means that our resellers can assure their customers that they will receive the same protections as customers who purchase qualifying Copilot offerings directly from Microsoft.

Since the launch of our AI Customer Commitments, we have engaged in hundreds of meetings and events with customers globally, providing insights into how they can leverage Copilot and Azure OpenAI in their day-to-day work and helping them evaluate and address risks by leveraging our responsible AI resources. We have created a variety of resources for customers, ranging from guidance on prompt engineering for lawyers to answers to frequently asked questions on data privacy and security. We also continue to help customers understand how to comply with the myriad of new regulations, such as the newly enacted EU AI Act.

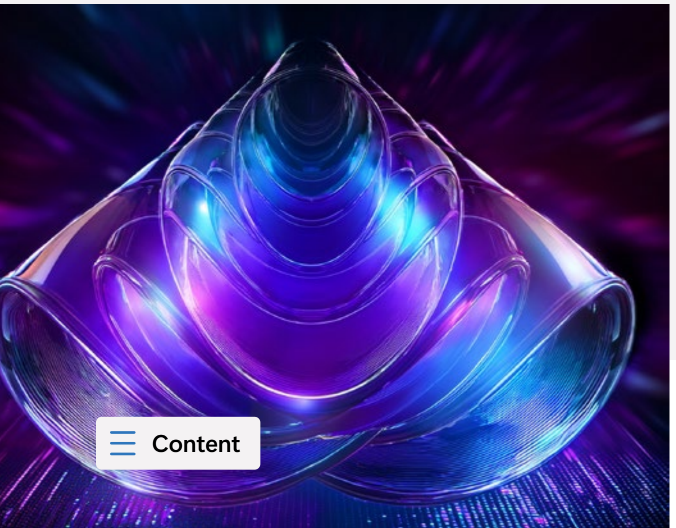
One of the core concepts of the EU AI Act is that obligations need to be allocated across the AI supply chain. This means that upstream regulated actors, like Microsoft in our capacity as a provider of AI tools, services, and components, must support downstream regulated actors, like our enterprise customers, when they integrate a Microsoft offering into a high-risk application. We embrace this concept of shared responsibility and aim to support our customers with their AI development and deployment activities by sharing our knowledge, providing documentation, and offering tooling, which we explore in greater detail below.

Provisions of the EU AI Act take effect on a staggered timeline over the next several years. As such, Microsoft is prioritizing regulatory compliance with the provisions that take effect first. We are publishing documentation and resources related to the EU AI Act on the Microsoft Trust Center on a regular basis to provide updates and address customer questions.[60] We also regularly update our Responsible AI Resources site, a rich source of tools, practices, templates, and information intended to help customers establish the foundations of good governance to support EU AI Act compliance.[61]

Read more about the EU AI Act on page 10 >

Tooling to support responsible deployment

Responsible AI tooling is critical to our own efforts to achieve consistent and efficient alignment with our internal AI policies and prepare for compliance with regulatory requirements. To empower our customers to do the same, we've released a total of 30 responsible AI tools that include more than 155 features to support their responsible AI development. Forty-two of these 155 features were released in 2024. These include automated red teaming capabilities available in Azure AI Foundry; new measurement and mitigation capabilities in Azure AI Foundry through offerings such as Azure AI Content Safety and Azure AI Evaluation SDK; and new AI security and privacy features in Microsoft Purview and Microsoft Defender.



Tools to map and measure risks

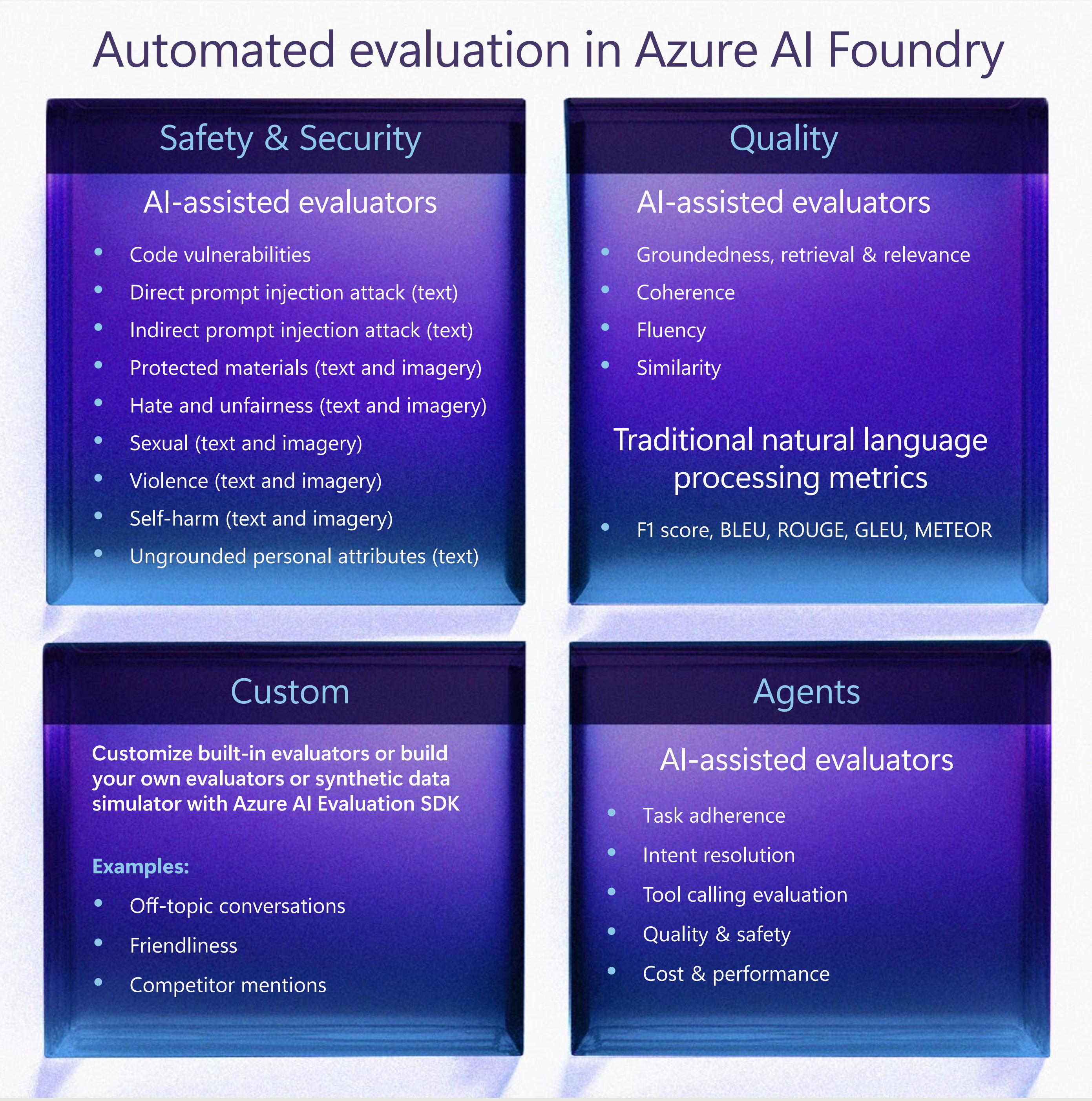
We continue to update our risk mapping and measurement tools and release new ones based on our own internal learnings, customer feedback, and constantly evolving technology.

In 2024 and early 2025, we launched new capabilities in Azure AI Foundry to help organizations to simplify model selection and map a broader range of potential risks in their AI models and systems.[62] We introduced model leaderboards in Azure AI Foundry, which allow users to compare models by quality, cost, and performance, explore trade-offs, and access detailed benchmarks.[63] To support automated red teaming, we announced PyRIT's integration with Azure AI Foundry, which allows our customers to simulate adversarial attack techniques and generate red teaming reports directly in Azure AI Foundry.

We also released new capabilities to measure specific safety and security risks such as the risk of reproducing protected materials, code vulnerabilities for AI systems that generate code, and indirect prompt injection attacks.[64] Such attacks, also known as XPIA, are an emerging attack vector where a threat actor poisons a model's grounding data source—such as a public website, email, or internal document—to pass hidden, malicious instructions to a model and circumvent safety and security guardrails.

To support safe deployment of agentic AI systems, we released new evaluation capabilities in Azure AI Foundry.[65] Agentic AI evaluation capabilities in Azure AI Foundry include intent resolution, tool calling accuracy, task adherence, and more.

Lastly, to enable continuous monitoring of AI systems in production, we released Azure AI Foundry Observability in May 2025. Integrated with Azure Monitor Application Insights, Azure AI Foundry Observability offers continuous monitoring and evaluation of both AI applications and agentic AI systems in production.[66]



Automated evaluation in Azure AI Foundry

Safety & Security

AI-assisted evaluators

- Code vulnerabilities
- Direct prompt injection attack (text)
- Indirect prompt injection attack (text)
- Protected materials (text and imagery)
- Hate and unfairness (text and imagery)
- Sexual (text and imagery)
- Violence (text and imagery)
- Self-harm (text and imagery)
- Ungrounded personal attributes (text)

Quality

AI-assisted evaluators

- Groundedness, retrieval & relevance
- Coherence
- Fluency
- Similarity

Traditional natural language processing metrics

- F1 score, BLEU, ROUGE, GLEU, METEOR

Custom

Customize built-in evaluators or build your own evaluators or synthetic data simulator with Azure AI Evaluation SDK

Examples:

- Off-topic conversations
- Friendliness
- Competitor mentions

Agents

AI-assisted evaluators

- Task adherence
- Intent resolution
- Tool calling evaluation
- Quality & safety
- Cost & performance

Tools to manage risks

In 2024, we released several new risk mitigation capabilities in Azure AI Content Safety for our customers to leverage as they build their own AI systems in Azure. These new mitigations expand risk mitigation support for multiple modalities, offer new deployment options, allow our customers to correct ungrounded content in real-time, and build mitigations for custom risk categories.

One of the risks that can compromise both accuracy and trustworthiness in an AI system is its propensity to generate content that does not align with information from input sources, also known as ungrounded content. In March 2024, we introduced groundedness detection, a feature that identifies and filters ungrounded content in AI outputs, helping developers enhance the accuracy and usefulness of generative AI applications by pinpointing responses that lack a foundation in connected data sources. In September, we launched a preview of "correction," a new capability within Azure AI Content Safety's groundedness detection feature, for our standalone API offering.[67] The correction capability corrects ungrounded AI outputs in real-time before users encounter them.

To use groundedness detection, a generative AI application must connect to grounding documents or input sources from which to base its outputs. Once the correction capability is enabled, the detection of an ungrounded sentence triggers a new request to compare the flagged generated output against the grounding documents. The system will then either remove the sentence if no relevant content exists in the grounding documents or rewrite the sentence to better align it with the grounding documents.

Also in September 2024, we launched our multimodal model in public preview in Azure AI Content Safety.[68] The multimodal model analyzes materials containing both image and text content to help our customer make applications and services safer from harmful user-generated or AI-generated content.

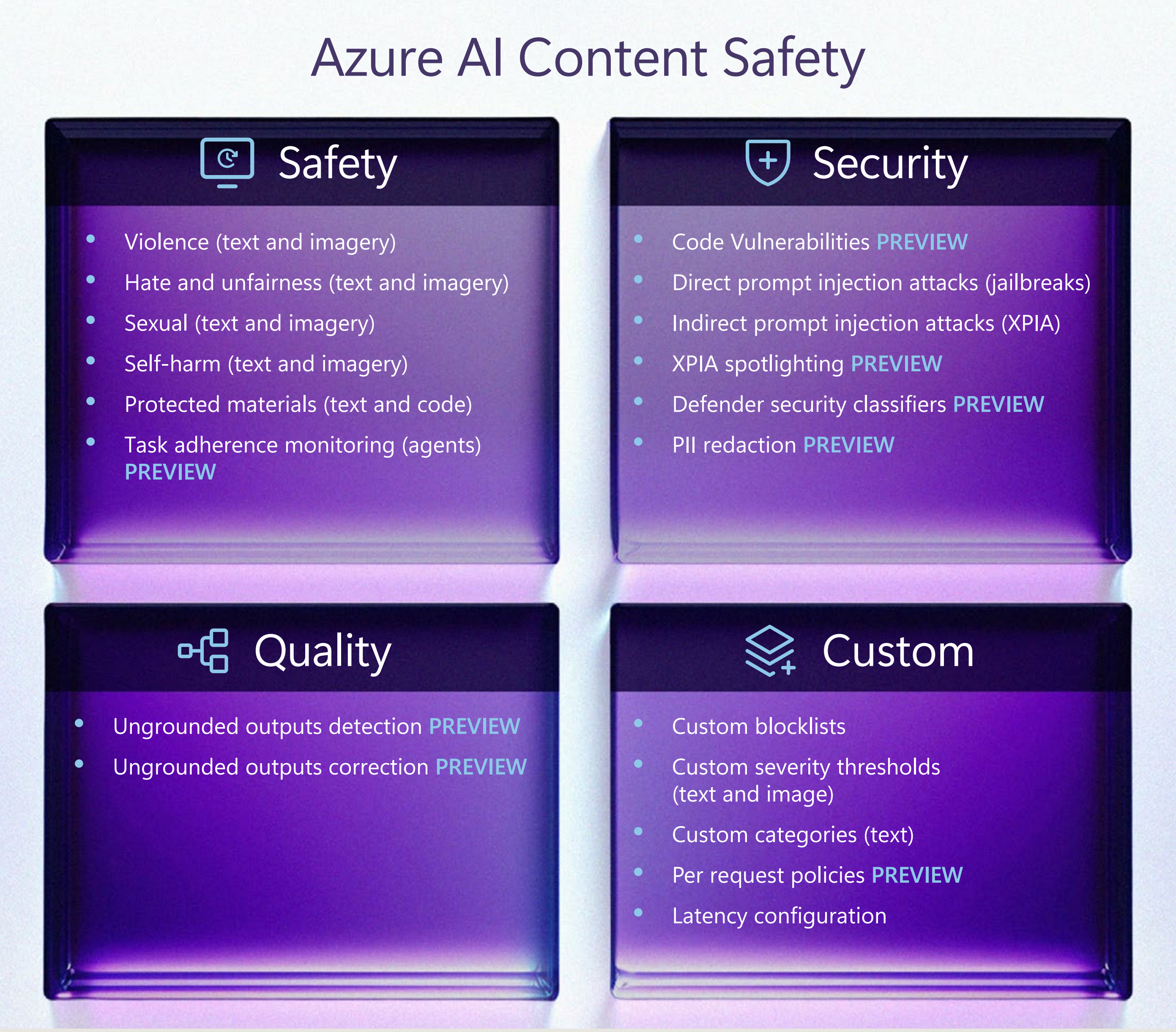
Key objectives of the multimodal model include:

- **Detecting harmful content across multiple modalities:** Analyzing both text and images (including emojis) to detect sexual, violent, and self-harm content and content related to hate and unfairness within text-image combinations.
- **Conducting contextual analysis across text and images:** The combined elements of text and images may convey richer meaning than either mode alone. The multimodal model can analyze text and images that occur together to detect subtle or implicit harmful content that might not be evident when looking at the text or image in isolation.
- **Moderating content in real time:** The model also provides real-time detection and moderation to help our customers prevent the generation, sharing, or dissemination of harmful content across multimodal platforms as appropriate for their use case. This ensures our customers have the tools they need to address harmful content before it reaches users.

In October 2024, we released the custom categories feature as part of Azure AI Content Safety, which allows customers to create and manage their own content moderation categories with two different APIs, depending on desired functionality. The custom categories standard API enables customers to define categories specific to their needs, provide sample data, train a custom machine learning model, and use it to classify new content according to the learned categories. The custom categories rapid API is designed to be quicker and more flexible than the standard method. It uses an LLM to quickly learn specific content patterns, allowing customers to react quickly and remediate incidents that are important to their use case.

In 2024, we also released Embedded Content Safety, which allows customers to embed Azure AI Content Safety on devices.[69] This capability is important both for on-device scenarios where cloud connectivity might be intermittent or unavailable and situations where companies don't want cloud connectivity for security or privacy reasons. These features scan text or image content for sexual, violent, and self-harm content and content related to hate and unfairness across multiple severity levels.

In May 2025, Azure AI Content Safety introduced new security classifiers from Microsoft Defender, and capabilities to mitigate code vulnerabilities, redact personally identifiable information, monitor task adherence for agentic AI systems and more flexibility for customers to customize their own mitigations.



Azure AI Content Safety

Safety

- Violence (text and imagery)
- Hate and unfairness (text and imagery)
- Sexual (text and imagery)
- Self-harm (text and imagery)
- Protected materials (text and code)
- Task adherence monitoring (agents) PREVIEW

Security

- Code Vulnerabilities PREVIEW
- Direct prompt injection attacks (jailbreaks)
- Indirect prompt injection attacks (XPIA)
- XPIA spotlighting PREVIEW
- Defender security classifiers PREVIEW
- PII redaction PREVIEW

Quality

- Ungrounded outputs detection PREVIEW
- Ungrounded outputs correction PREVIEW

Custom

- Custom blocklists
- Custom severity thresholds (text and image)
- Custom categories (text)
- Per request policies PREVIEW
- Latency configuration

Pioneering safe generative AI in South Australian classrooms with Azure AI Content Safety

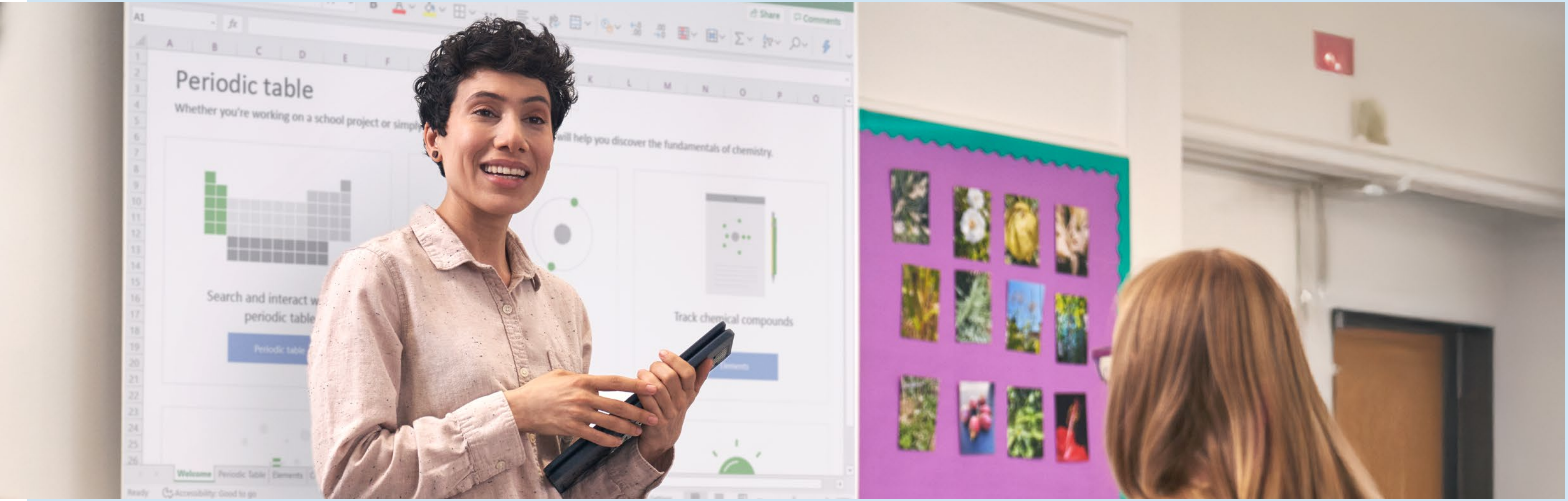
Recognizing the need to adapt to rapid advancements in AI, the South Australia Department for Education sought to embrace the transformative potential of generative AI technologies in the classroom. Before bringing generative AI to students and teachers, they needed to address the critical question of how to do it responsibly.[70]

Protecting students from potentially harmful or inappropriate content was a core concern for the South Australia Department for Education. In collaboration with Microsoft, the Department tackled these concerns by launching "EdChat," an educational chatbot with built-in safeguards. To ensure a safe experience with EdChat, the Department implemented Microsoft's Azure AI Content Safety, which employs advanced models to identify and mitigate harmful and risky content.

This safeguarding measure was crucial for the successful deployment of EdChat in schools. Azure AI Content Safety provided EdChat with built-in measures and controls to detect and block potentially harmful AI-generated content from reaching students. The Department retained full control over how Azure AI Content Safety was configured to detect and block potentially harmful content in EdChat.

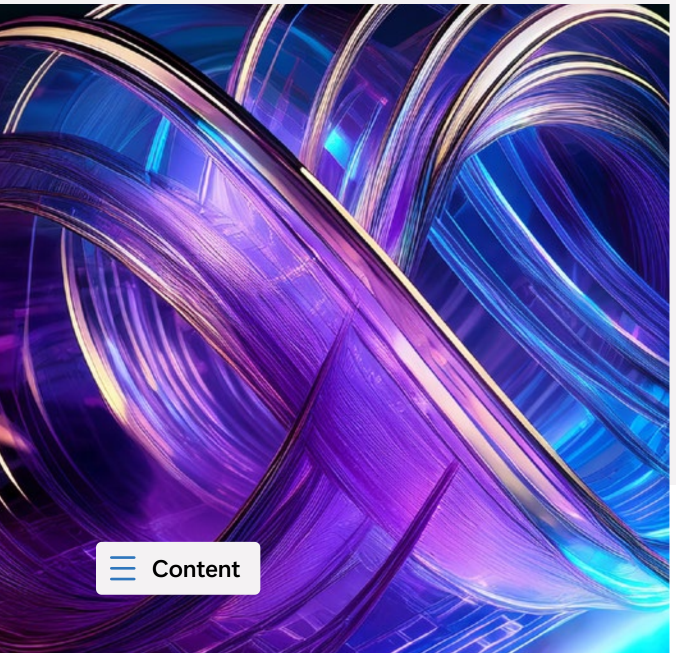
The Department launched a pilot program to gauge the effectiveness of EdChat, involving approximately 1,500 students and 150 teachers from eight secondary schools over an initial eight-week trial period. This trial aimed to explore the chatbot's capabilities in assisting with research and enhancing educational experiences. School principals were given autonomy to decide the extent of EdChat's usage among students and teachers, tailoring the experience to their school's specific needs.

During the trial phase, approximately 20% of students were actively using EdChat. This number increased significantly as both teachers and students became more familiar with the tool's potential to support critical and creative thinking. The built-in safety features to block inappropriate queries and filter harmful responses allowed teachers to focus more on the technology's educational benefits than content oversight. The successful trial solidified the Department's confidence in EdChat's underlying architecture and the effectiveness of its safety measures.

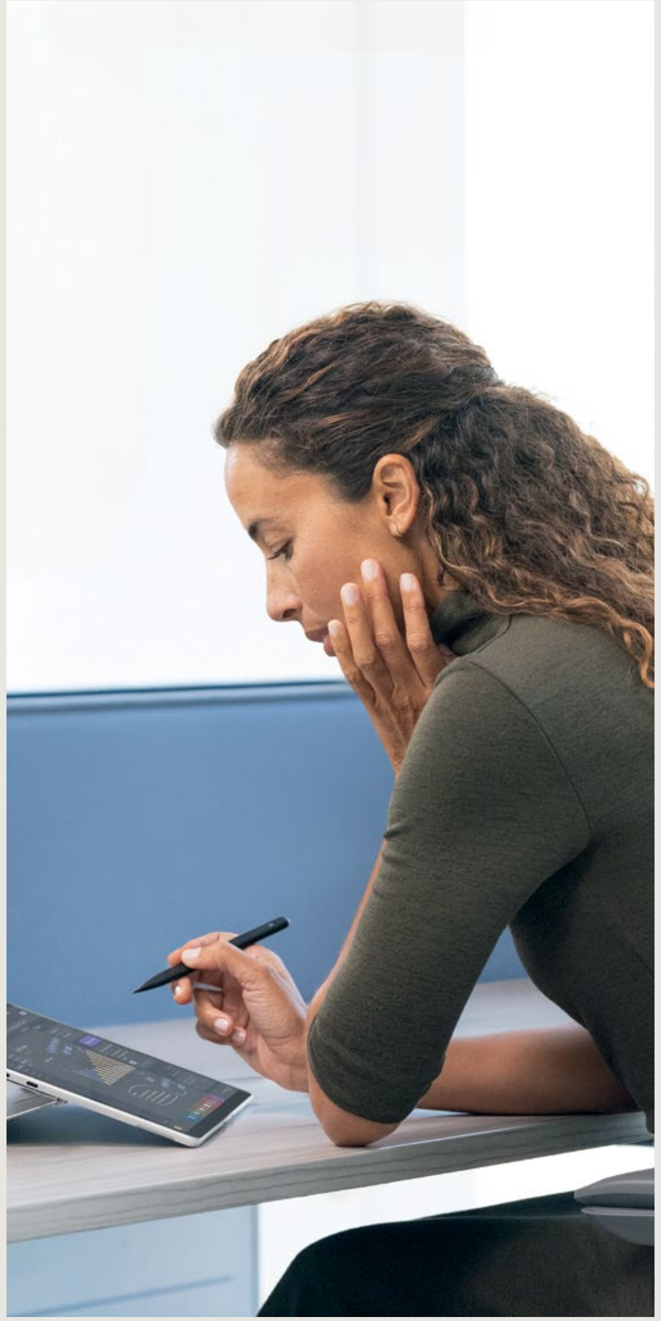

Periodic table
Whether you're working on a school project or simpl
will help you discover the fundamentals of chemistry.
Search and interact w
periodic table
Track chemical compounds

Transparency to support responsible development and use by our customers

As a provider of AI tools, services, and components, we understand that we must do our part to equip our customers with the information they need to innovate responsibly. This is particularly true for our platform services like Azure OpenAI Service and Azure AI Search, which offer our customers powerful AI capabilities to build their own AI models or applications. Since 2019, we've published 40 Transparency Notes, all of which contain key information about the capabilities, limitations, and intended uses of these platform services.



In 2023, we expanded our transparency documentation to require our non-platform services, such as our Copilots, to publish similar information in Responsible AI Frequently Asked Questions (FAQs) and other important disclosures. This includes, for example, in-product disclosure in products like Microsoft Copilot and M365 Copilot to inform users they're interacting with an AI application, as well as citations to source material.



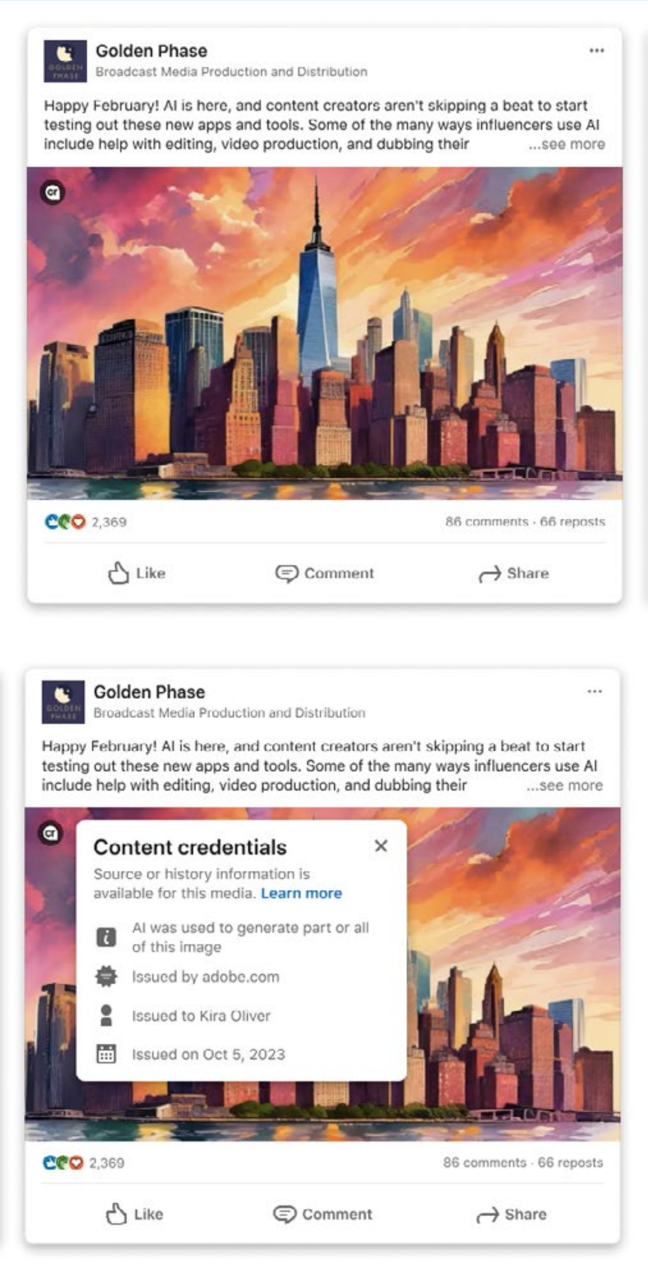
Content Credentials on LinkedIn

In 2024, Microsoft-owned platform LinkedIn became the first professional networking platform to implement direct disclosure of content provenance.[71] Our aim in launching this feature was both to provide LinkedIn users with the proper tools to formulate their own opinion on whether to trust the media content they see and to contribute to industry adoption of direct disclosure of content origin and history.

In May 2024, LinkedIn began displaying the C2PA Content Credentials—metadata that includes key information about the provenance of the content—next to a subset of AI-generated images and videos.[72] Available metadata for these images and videos included:

- **Assertion about the media:** Information about whether AI was used to generate all or part of the content, an assertion that comes from the entity that created and signed the content credential.
- **Content source:** The origin of the content, such as the camera model or AI tool used to generate all or part of the image.
- **Issued by:** The entity that created and signed the content credential, which could be an individual creator, an organization, or a trusted authority.
- **Issued on:** The date and time the Content Credential was created and signed.

After user feedback indicated that incorporating more information would be helpful, the team continued iterating. The second phase of the release displayed Content Credentials for all images and videos uploaded to LinkedIn's feed with cryptographically signed metadata. While the first iteration noted if AI was used to generate part or all of an image, the second iteration included more granular details for synthetic content, differentiating between assertions that (1) AI was used to generate part of an image, (2) AI was used to generate all of an image, or (3) AI was used to edit an image. By continuing to expand and improve direct disclosure mechanisms, LinkedIn aims to help accelerate industry adoption and user education around these mechanisms.



Golden Phase
Broadcast Media Production and Distribution

Happy February! AI is here, and content creators aren't skipping a beat to start testing out these new apps and tools. Some of the many ways influencers use AI include help with editing, video production, and dubbing their ...see more

Golden Phase
Broadcast Media Production and Distribution

Happy February! AI is here, and content creators aren't skipping a beat to start testing out these new apps and tools. Some of the many ways influencers use AI include help with editing, video production, and dubbing their ...see more

Content credentials

Source or history information is available for this media. Learn more

AI was used to generate part or all of this image

Issued by adobe.com

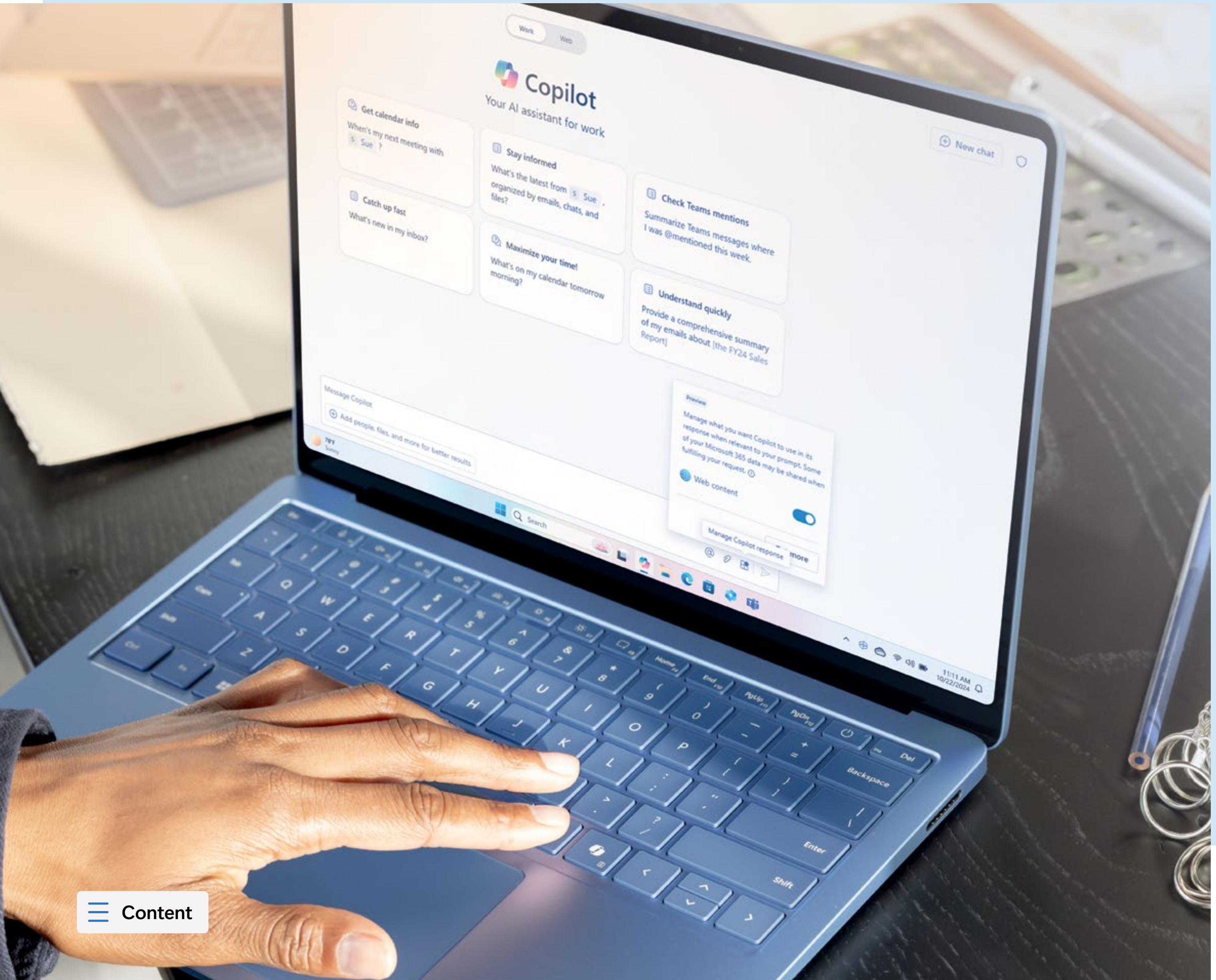
Issued to Kira Oliver

Issued on Oct 5, 2023

Greater transparency for search queries in Copilot

To receive the highest quality responses from Copilot and M365 Copilot, users have the option to allow Copilot to reference web content. In September 2024, in response to customer interest in increased transparency into the source material scanned by Copilot, we announced new features that allow both users and admins greater visibility into Copilot-generated web queries.[73] Providing transparency into web queries furthers our goal of advancing trustworthy AI by allowing customers to confirm that the information searched by Copilot is relevant and appropriate.

Since this update, web search query citations (generally available for Microsoft 365 Copilot Business Chat and Microsoft Copilot) now include the exact web search queries derived from the user's prompt. This information, which can be found in the linked citation section of the Copilot response, allows users to understand what search queries and sites were used to enhance Copilot's answer. Web search query logging (generally available for Microsoft 365 Copilot Business Chat) enables admins to perform search, audit, and eDiscovery on the exact web search queries Copilot derived from the user's prompt. Admins, who were previously able to perform these actions for prompts and responses, can now extend those actions to search queries.



AI literacy efforts

Promoting AI literacy helps ensure that individuals can make informed decisions regarding AI systems. The EU AI Act requires that deployers and providers of AI systems "provide all relevant actors in the AI value chain with the insights required to ensure the appropriate compliance and its correct enforcement."[74] The AI Act does not require a one-size-fits-all literacy program—rather, it leaves room for each provider and deployer to take into account the training and experience of their personnel, the context of the AI systems they use, and how AI systems will be used.

Microsoft has an extensive knowledge base of AI literacy materials developed over years of experience in the field of AI. We offer a variety of AI literacy programs for our employees, with role-specific online learning paths tailored to their technical expertise and the ways in which they develop, provide, and use AI systems. Microsoft also integrates AI education into our annual Trust Code (Standards of Business Conduct) to ensure that all employees understand our Responsible AI Standard.

We provide a range of public resources to help increase AI literacy for organizations and individuals, including AI courses that range from basic education on what AI is to technical training on responsible AI development and use. Our Office of Responsible AI released an AI Literacy Starting Guide with training materials for vendors and customers targeted at different roles and levels of knowledge.[75] Microsoft Philanthropies also created an AI Skills Navigator, which leverages generative AI technology to enable learners to create personalized AI learning pathways based on their background and experience.[76] Lastly, in July 2024, the Older Adults Technology Services (OATS) from AARP announced plans to develop and deploy a training program to further AI education and literacy among American adults over the age of 50. This program will be funded by the Societal Resilience Fund, established by both OpenAI and Microsoft.[77]

While customers need to assess how literacy requirements in the EU AI Act and other regulatory efforts apply to them in the specific context of their organization, our AI literacy materials provide a useful starting point. As Microsoft and our customers continue to innovate with AI, we are committed to sharing our learnings and new materials so that every person and organization has the resources they need to build and improve their AI literacy.

Section 4

How we learn, evolve, and grow

From the beginning, Microsoft has committed to scaling our responsible AI program to meet the growing demand for this technology. In this section, we provide highlights from 2024 on our work to learn, evolve, and grow. For us, this means investing in research, working across sectors to advance effective global governance of AI, and tuning into a wide range of perspectives.

Investments in research

Microsoft researchers work to advance the state of the science of responsible AI with the aim of enhancing our understanding of AI, creating new model architectures with novel capabilities, achieving societal benefit, transforming scientific discovery, and extending human capabilities. In 2023, Microsoft formed the AI & Society research network, which addresses the many bidirectional relationships between AI technologies and people, groups, organizations, and society as a whole. The network aims to bring together a diverse and multidisciplinary community of researchers to explore and shape the social and technical aspects of AI. In addition to pursuing the aims above, the research network explores topics that include sociotechnical evaluation and alignment of AI and equitable AI.

As a complement to the efforts of this research network, Microsoft launched the Microsoft Research AI & Society Fellows program[78] in 2023 to catalyze collaboration between Microsoft researchers and eminent scholars and experts across a range of disciplines at the intersection of AI and societal impact. This investment has resulted in 13 distinct, ambitious research collaborations bringing together Microsoft researchers and 24 esteemed fellows across academic and industrial disciplines. These research collaborations address some of the pressing challenges facing organizations that aim to advance responsible AI practices, including research focused on advancing the science of AI risk measurement, as described in the Measurement section earlier in this report.

Read more about AI risk measurement research on page 18 ›

Other research collaborations within the AI & Society Fellows program have a broader focus. For example, one of the research collaborations aims to address how to most effectively regulate AI in light of the challenges of doing responsible AI in practice. The collaboration aims to take stock of existing social scientific insights into the difficulties faced by regulated entities seeking to comply with existing and forthcoming regulations, complete original empirical studies to fill identified gaps in the existing scholarship on responsible AI in practice, and channel these findings into the ongoing debates about how to craft effective regulations of AI.

Microsoft also established the AI Frontiers lab in 2024 to invest in the core technologies that push the frontier of what AI systems can do in terms of capability, reliability, and efficiency.[79] Researchers at our AI Frontiers lab are not only interested in how well these systems work—they also want to ensure that we build in sociotechnical solutions that can make these systems work in a responsible way.

Our research teams work in close collaboration with our policy and engineering teams to inform our approach to responsible AI. Throughout 2024, Microsoft researchers pushed the frontiers of our understanding of mapping, measuring, and managing AI risks. We summarize some of their research contributions on the following pages.

Advancing the science of mapping risks

Advancements in generative AI systems have enabled the development of agentic AI systems that are capable of autonomously executing actions and collaborating with other agents to achieve user-specified goals. Building on prior research, researchers across Microsoft are continuing to support our collective understanding of the emerging risks associated with the development and use of agentic AI systems, including the wide range of failure modes with human-agent communication.[80]

Over the past year, Microsoft researchers, in collaboration with a broader set of researchers across academia and the technology industry, contributed new approaches to advance existing responsible AI tools and practices, such as impact assessments and AI documentation frameworks. This includes a study that explores the use of impact assessments by industry researchers and contributes 10 design considerations to facilitate the effective design, development, and adaptation of an impact assessment template for use in industry research settings and beyond.[81] In May 2024, researchers introduced the CLeAR (Comparable, Legible, Actionable, and Robust) framework, which aims to help practitioners consider the complexities and tradeoffs required when developing documentation for datasets, models, and AI systems throughout their lifecycles.[82]

Researchers also probed the potential risks associated with the use of conversational AI systems for social and emotional support. Through this study, researchers developed a taxonomy and framework that advance our understanding of AI behaviors, psychological impacts, and the contexts in which these impacts may manifest. They recommend emphasizing AI interaction disclosure, specifically focusing on stressing the non-human nature of the system, and when sensitive topics are discussed, enabling the system to gracefully redirect the user to accessible and actionable resources.[83]

These research insights serve as critical inputs to Microsoft's AI policy development and oversight efforts. For example, the Sensitive Uses and Emerging Technologies team collaborates closely with researchers studying the risks of using conversational AI systems for emotional support, which informs how the team crafts early-steer guidance aimed at providing engineering teams with actionable recommendations.

Advancing the science of measuring risks

Researchers across industry and academia are uniquely positioned to make meaningful contributions to advance the science of AI risk measurement. Within Microsoft Research, the Sociotechnical Alignment Center (STAC) continues to produce meaningful research and thought leadership in this area.[84] Building on the four-part measurement framework discussed earlier in the report, the STAC team has published papers that extend this framework to include the systematization and operationalization of amounts, populations, and instances,[85] as well as a set of general dimensions that capture critical choices involved in generative AI evaluation design.[86] STAC has also published work to help bridge risk mapping and measurement by investigating whether red teaming can produce measurements that enable meaningful comparisons of systems.[87] These research contributions advance our ability to better understand, interrogate, and compare different evaluation methods.

Other Microsoft Research teams made additional contributions that advanced the science of AI risk evaluation in 2024, ranging from studies that focus on the use of synthetic data, including in AI evaluation tasks,[88] releasing open-source tools such as practical evaluation tools to assess the safety of AI systems that can execute code.[89] As highlighted earlier in the report, the AI Frontiers lab provided meaningful thought leadership and scalable tools to fill gaps in benchmark evaluation practices across the industry.

Advancing the science of managing risks

Researchers across Microsoft continue to explore novel strategies to manage the risks associated with the use and misuse of AI. These include mechanisms to protect against indirect prompt injection attacks, reduce the risk of overreliance on AI outputs,[90] and develop novel approaches to steer model behavior.

Identifying novel strategies to defend against indirect prompt injection attacks (XPIA), an attack mechanism where threat actors embed hidden malicious instructions in a grounding data source to circumvent safety guardrails, continues to be an area of interest for both practice and research. In 2025, Microsoft researchers, in collaboration with researchers in academia, developed a benchmark to assess the risk of XPIA vulnerabilities in LLMs, now available on GitHub.[91] In their research, they also identified promising defense mechanisms to defend against XPIA, including boundary awareness to help LLMs differentiate between user prompts and external content, and explicit reminders for the LLM not to execute instructions embedded in external content.[92]

Exploring ways to reduce inappropriate reliance on AI-generated output was a recurrent theme in research throughout 2024. Researchers across Microsoft contributed to a growing body of work on fostering appropriate reliance. One study explored the impact of LLM's uncertainty expression, comparing first-person with general perspective.[93] Another paper explored the risk of overreliance on generative AI and identified emerging mitigation techniques such as uncertainty highlighting, cognitive forcing functions, contrastive explanations, and AI critiques.[94]

Researchers also explored how to steer model behavior towards safer and more reliable outputs while preserving performance. In one study, researchers explored inference time interventions where they first identified features that mediated refusals and then assessed whether amplifying these features improved robustness to challenging multi-turn jailbreak attacks while preserving model performance.[95]

Advancing AI research beyond Microsoft

In addition to conducting our own research, we also support academic research by providing key resources for AI research. For example, last year, we announced our support of the National AI Research Resource (NAIRR) pilot led by the National Science Foundation, which provides high-quality data, computational resources, and educational support to make cutting-edge AI research possible for more U.S. academic institutions and non-profits. Microsoft has committed $20 million worth of Azure compute credits to support researchers by providing high-performance computing resources and access to leading-edge models.[96] Our commitment to the NAIRR pilot also includes collaborative opportunities with Microsoft's scientists and engineers as well as resources to accelerate domain-specific research such as innovative tools for chemistry and materials science research via Azure Quantum Elements[97] and tools for research and development on AI fairness, accuracy, reliability, and interpretability.[98]

To date, we have provided 38 grants for researchers across the U.S. to access critical resources for AI research and development on Azure. These grants support research projects at both academic institutions and non-profit organizations focused on molecular biology and protein design, healthcare and drug discovery, sustainability and earth sciences, personalized education and accessibility, agriculture, human-AI collaboration, and new approaches for AI privacy, safety, and security.

Foundation models continue to fuel a fundamental shift in computing research, natural sciences, social sciences, and computing education itself. The Accelerating Foundation Models Research (AFMR) initiative was created by Microsoft Research to work with the broader academic research community to enable AI advances and nurture a vibrant and diverse AI research ecosystem by providing access to state-of-the-art foundation models hosted on Microsoft Azure. The goal of AFMR is to foster increased collaboration across disciplines, institutions, and sectors to unleash the full potential of AI for a wide range of research questions, applications, and societal contexts.

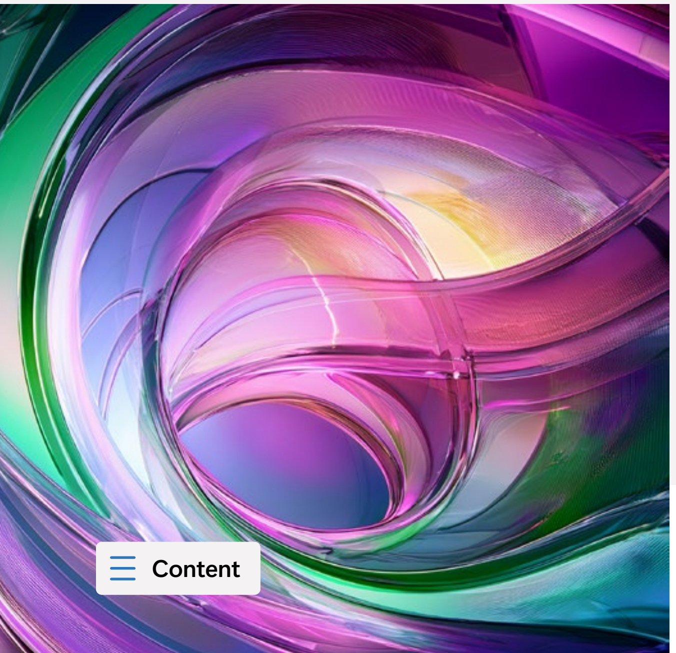
To date, the AFMR research community has published over 300 papers co-authored by computer scientists and researchers outside computer science supporting over 123 institutions in 19 countries.[99] This depth and breadth of expertise across disciplines, cultures, and languages has contributed meaningfully to our collective ability to use AI to address some of the world's greatest challenges around the following three goals:

- **Aligning AI with shared human goals, values, and preferences via research on models** to enhance safety, robustness, sustainability, responsibility, and transparency, while also exploring new evaluation methods to measure the rapidly growing capabilities of new models.
- **Improving human-AI interactions via sociotechnical research**, which enables AI to extend human ingenuity, creativity, and productivity, while also working to reduce inequities of access and working to ensure positive benefits for people and societies worldwide.
- **Accelerating scientific discovery in natural sciences** through proactive knowledge discovery, hypothesis generation, and multiscale multimodal data generation.

Working together as a global research community is essential to realizing the promise of AI to benefit each individual, organization, and society as a whole. AFMR is one means by which we make progress towards these goals.

Advancing AI innovation and adoption through good governance

Across jurisdictions, horizontal and issue-specific AI laws, norms, and standards are advancing at the same time. Horizontal approaches, such as the European Union's AI Act, address multiple layers of the technology stack, multiple sectors, and multiple issues, while narrower approaches focus on specific governance measures or topics, like synthetic media or frontier model safety. Microsoft engages in global efforts to build consensus-based frameworks, promoting coherence across borders while instilling clear allocation of responsibilities across the AI value chain.



Previous waves of technology have demonstrated that there are two components to the trust that underpins broad adoption and iterative innovation: first, trust in how technology itself performs; and second, confidence that people and organizations can deploy it successfully. In order to continue to advance trust and confidence in AI, we must work towards globally coherent governance frameworks that can help accelerate adoption and allow organizations of all kinds to innovate and use AI across borders. Microsoft will continue to share the lessons learned from our internal governance work so that others may build on it, focusing on three key areas:

- **Strengthening feedback loops between innovation and governance.** We know even greater AI capability is on the horizon, with the opportunity to unlock innovation in science, education, and countless other fields. AI deployment, experimentation, and skilling must go hand-in-hand with AI governance to create tighter feedback loops on what is effective in practice. As we've learned through implementing Microsoft's Responsible AI Standard, while minimum guardrails provide an important starting point, we can learn much more about how to govern AI technology effectively in practice through AI deployment and experimentation. Moreover, as we've learned through establishing Microsoft's responsible AI program and developing role-based training, good governance includes investing in people so that they can take advantage of the capabilities that AI already demonstrates and strengthen their readiness for the AI capabilities that may emerge. If we get this technology into the hands of more people who can apply it to the local challenges that they uniquely understand, then we will not only have much greater impact in realizing opportunities but also a much broader feedback loop on governance.
- **Advancing scientific understanding to inform effective guardrails and practice.** Over the past two years, stakeholders around the world have come together to define high-level objectives for AI governance. As work towards interoperable governance across borders continues, we are also turning our attention towards building a deeper shared understanding of effective and easy to adopt risk management techniques and technical practices that can help realize the high-level goals of these governance frameworks. At Microsoft, we are continuing to invest in our own internal governance frameworks, learning from their implementation while also participating in multistakeholder efforts. Partnering with government research bodies, such as AI institutes and centers for standards and innovation, and collaborating on consensus-driving publications like the International AI Safety Report can help us close identified evidence gaps and synthesize research and applied learnings. Frameworks for voluntary reporting on governance practices, such as the Hiroshima AI Process Reporting Framework, can also help to deepen, align, and streamline shared expectations across jurisdictions.

- **Aligning expectations for guardrail implementation.** Today's AI systems often involve multiple models and components from different providers. For organizations deploying these systems to have confidence in each component and the system as a whole, it's important to align on expectations for guardrail implementation across different actors in the supply chain. Building on developing industry norms, as Microsoft has done through our Frontier Governance Framework[100] to address national security risks of highly capable models, will help accelerate progress. Where expectations for the behavior of AI applications diverge among jurisdictions, focusing on where guardrails can align—such as on expectations for model-level transparency—will yield significant benefits for AI adoption and innovation.

To make further progress in advancing AI governance, we must develop ecosystem-wide reference points for effective guardrails as well as tools that support implementation for ourselves and our customers. Just as internal investments in cybersecurity practices and tools have enabled us to support the broader cybersecurity ecosystem, so too have our years building an AI governance program readied us to help others seize AI opportunities. We will continue to make such investments and work through organizations like the Frontier Model Forum and MLCommons to develop industry practice reference points and AI evaluation tools.

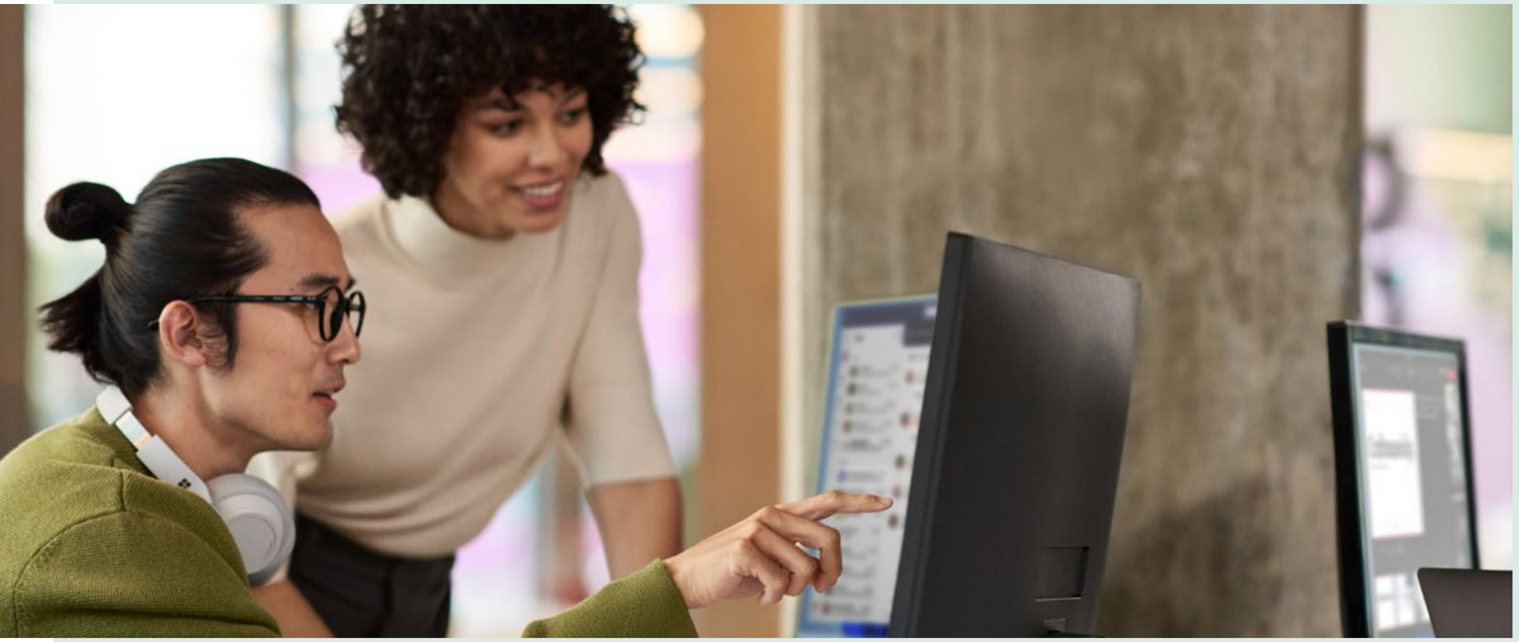
Case study: AILuminate from ML Commons

Developers and deployers of AI technologies have a shared interest in developing transparent and practical safety assessments to guide development decisions, inform purchasers and consumers, and support standards bodies and policymakers. MLCommons leveraged this momentum to develop a new AI safety benchmark called AILuminate. To support this effort, technical experts from Microsoft are part of the MLCommons AI Risk and Reliability working group—a team of leading AI researchers from institutions including Stanford University, Columbia University, and TU Eindhoven; civil society representatives; and technical experts from Google, Intel, NVIDIA, Meta, Microsoft, Qualcomm Technologies, Inc., and others.[101]

In December 2024, MLCommons released AILuminate, a first-of-its-kind safety test for large language models (LLMs) that represents important progress in developing research-backed, effective evaluation techniques for AI safety testing.[102] The v1.0 benchmark provides a series of safety grades for the most widely used LLMs and offers a scientific, independent analysis of LLM risk that can be immediately incorporated into company decision-making.

The AILuminate v1.0 benchmark assesses the safety of text-to-text interactions with a general purpose, English-language AI chat model by a naive or moderately knowledgeable user with malicious or vulnerable intent. The MLCommons AILuminate benchmark evaluates an AI system-under-test by inputting a set of 24,000 test prompts across twelve categories of hazards. These responses are recorded and then evaluated using a specialized set of "safety evaluator models" to determine which of the responses are violations of the AILuminate Assessment Standard guidelines.

During initial testing of a range of popular AI systems, none of the LLMs evaluated were given any advance knowledge of the evaluation prompts or access to the evaluator model used to assess responses. Microsoft's Phi 3.5 MoE Instruct (API) model received a grade of "very good" on a five-point scale in which "very good" represents a four out of five. The development of this benchmark is a critical step towards the formulation of industry-standard testing to guide responsible development, and one that is set to continue to expand as MLCommons prepares to release versions in French, Chinese, and Hindi.



Working towards effective and interoperable governance

As AI is an international technology that is used across borders, it is especially important that governance frameworks across the world are interoperable and coherent. We continue to share lessons from our governance program around the world to help advance a coherent approach and build on the work we have done to date to advance interoperable governance. In May 2024, we published an e-book that features a series of case studies exploring the history and evolution of governance for technologies that are used across borders, including aviation and atomic energy.[103] Drawing on this deep, expert insight, we came away with three high-level takeaways for AI:

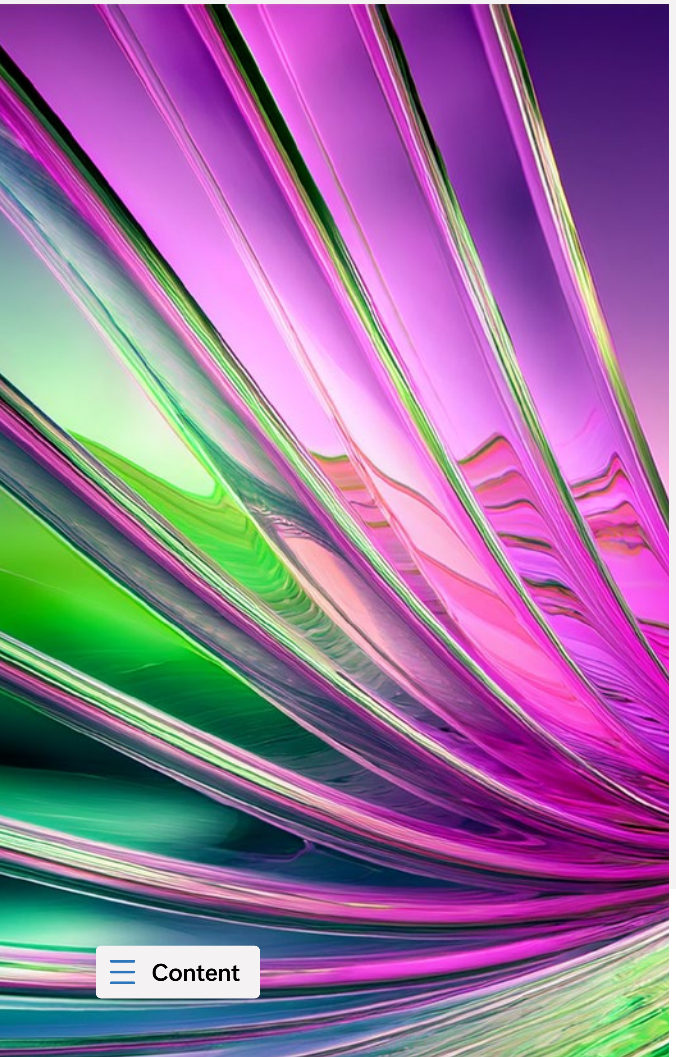
1. AI governance involves three interrelated layers: industry standards, domestic regulation, and international governance.
2. At the international governance layer, three outcomes are important for AI: globally significant risk governance, regulatory interoperability, and inclusive progress.
3. Four international governance functions will enable those outcomes: monitoring for and managing global risks, setting standards, building scientific consensus, and strengthening appropriate access to resources.

After publishing our global governance book, we led or participated in more than a dozen conversations with stakeholders around the world, including hosting events in Seoul, Tokyo, and São Luís, and joining events hosted by the Geneva Graduate Institute, the Eurasia Group in Toronto, the Tony Blair Institute and Centre for Governance of AI in London, Sciences Po and Renaissance Numérique in Paris, and Renaissance Numérique in Brussels.

During these conversations, stakeholders reinforced the importance of closing gaps between policy and science, as well as reaching consensus on how we define AI risk. In September 2024, we published an updated version of the global governance book that incorporated insights gleaned from these conversations. We look forward to continuing this dialogue in 2025 and beyond.

Tuning in to multistakeholder input

Harnessing the expertise of a wide range of stakeholders is essential to effective AI risk management research and practice. We actively seek input from a wide range of stakeholders on how our AI systems can be safer and more reliable. In this section we highlight two out of many such efforts.



Tuning in to global perspectives

To better inform our responsible AI practices and ensure they are inclusive of perspectives from around the world, including the Global South, we launched our Global Perspectives on Responsible AI Fellowship in partnership with the Stimson Center's Strategic Foresight Hub in 2023. Our first class of fellows consisted of 15 AI experts from across academia, civil society, and both the public and private sectors. The fellows received a stipend to participate in a five-part discussion series focused on advancing AI responsibly in the Global South and publish summaries of these discussions.[104] To wrap up the first iteration of the fellowship, we launched a series of essays and videos highlighting the fellows' work on AI in their respective regions.[105] We also published a paper in collaboration with two fellows from Latin America as well as the CAF Development Bank of Latin America and the Caribbean exploring how AI can help advance the UN's sustainable development goals in the region.[106]

The second iteration of the fellowship program kicked off in July 2024 with eleven fellows staying on and nine new fellows added, bringing our total cohort to 20 fellows. Our second cohort included representation from the Democratic Republic of Congo, Egypt, Thailand, Indonesia, Argentina, Turkey, and many more. With our second class of fellows, we increased our focus on generative AI and modified the structure of the fellowship to more intentionally center the voices of the fellows in our work within Microsoft and in international arenas where topics on global governance of AI are discussed. In addition to participating in a series of discussions on the benefits and risks of AI in the context of the Global South, the fellows also contributed to focus group sessions where they provided input on AI policy efforts led by Microsoft's Office of Responsible AI.

This included seeking feedback from the fellows on the second version of our Responsible AI Standard and the "Global Governance: Goals and Lessons for AI" book. Inspired by the focus group discussions hosted around the governance book, five fellows—from Thailand, the Democratic Republic of Congo, Kyrgyz Republic, Egypt, and India—are co-authoring a white paper on inclusive global governance of AI.

Collaborating with content creators

As part of our efforts to gather input from a wide range of stakeholders, we engage in and promote ongoing dialogue with creative professionals about how AI can be used responsibly and effectively to enhance their workflows and creative outputs.

In June 2024, we partnered with an art center in Paris, the Grand Palais Immersif, to exhibit art from over a dozen artists worldwide. The exhibit, titled "Artificial Dreams" showcased ways that AI can be used as a tool to aid creative expression.[107] Microsoft hosted an event for French MPs to experience the artwork and to hear from artists on how they integrated AI into the creative process.

We also attended the Busan International Film Festival (BIFF), the largest film festival in Asia, and the adjacent business-focused Asian Contents & Film Market (ACFM) in Busan, South Korea. We met with creators and cultural organizations to discuss how they use AI in their work and provided training and showcasing for filmmakers on how Copilot can assist in script development and addressing notes, gaining insights from film budgets and schedules, marketing a project, building great pitch decks, and visualizing a story.

In October 2024, we announced a partnership with UK-based youth arts charity National Youth Theatre to offer 45 free workshops aimed at equipping young people aged 18-30 with skills related to using generative AI responsibly in creative sectors.[108] The workshops, developed in accordance with our Responsible AI principles, covered topics ranging from how AI is currently being used in TV and film production to how new technologies can make storytelling more inclusive and accessible, and how AI can be used alongside other new extended reality technologies.

We continue to explore new opportunities to engage with creative communities around the world and develop best practices on how they can use AI responsibly to augment human creativity.

Section 5

Looking ahead

The progress we have made thus far and the progress we intend to make over the course of the next year would not be possible without the commitment of our employees around the world. They remain at the heart of our commitment to earn, build, and keep the trust of our customers and society more broadly. We are especially grateful for the talented members of our responsible AI community who continue to pioneer best practices to build trust into our technology from the beginning.

2024 ushered in a wide range of breakthroughs in AI, and waves of adoption across organizations big and small. These developments unlocked opportunities and learnings across a range of domains—from science, education, and healthcare to retail, manufacturing, and agriculture. They also refocused attention on effective AI governance in practice.

As AI innovation and adoption continue to advance, our core objective remains the same: earning the trust that we see as foundational to fostering broad and beneficial AI adoption around the world. As we continue that journey over the next year, we will focus on three areas to progress our steadfast commitment to AI governance while ensuring that our efforts are responsive to an ever-evolving landscape:

1. **Developing more flexible and agile risk management tools and practices, while fostering skills development to anticipate and adapt to advances in AI.** To ensure people and organizations around the world can leverage the transformative potential of AI, our ability to anticipate and manage the risks of AI must keep pace with AI innovation. This requires us to build tools and practices that can quickly adapt to advances in AI capabilities and the growing diversity of deployment scenarios that each have unique risk profiles. To do this, we will make greater investments in our systems of risk management to provide tools and practices for the most common risks across deployment scenarios, and also enable the sharing of test sets, mitigations, and other best practices across teams at Microsoft. We also need to continue investing in our employees, who are at the heart of our risk management efforts. This will include continued investment in training to better equip our employees with the skills they need to fully participate in building and deploying trustworthy AI technologies.

2. **Supporting effective governance across the AI supply chain.** Building, earning, and keeping trust in AI is a collaborative endeavor that requires model developers, app builders, and system users to each contribute to trustworthy design, development, and operations. AI regulations, including the EU AI Act, reflect this need for information to flow across supply chain actors. While we embrace this concept of shared responsibility at Microsoft, we also recognize that pinning down how responsibilities fit together is complex, especially in a fast-changing AI ecosystem. To help advance shared understanding of how this can work in practice, we're deepening our work internally and externally to clarify roles and expectations. Our internal policies and practices will evolve to better account for these interconnections. For example, our transparency documentation for AI systems will begin incorporating mechanisms that clearly delineate the roles of Microsoft from the customer and end user. We will also share lessons learned from these efforts to contribute to emerging best practices and standards that support smoother and more accountable collaboration among supply chain actors.

3. **Advancing a vibrant ecosystem through shared norms and effective tools, particularly for AI risk measurement and evaluation.** The science of AI risk measurement and evaluation is a growing but still nascent field. We are committed to supporting the maturation of this field by continuing to make investments within Microsoft, including in research that pushes the frontiers of AI risk measurement and evaluation and the tooling to operationalize it at scale. We remain committed to sharing our latest advancements in tooling and best practices with the broader ecosystem to support the advancement of shared norms and standards for AI risk measurement and evaluation.

We look forward to hearing your feedback on the progress we have made and opportunities to collaborate on all that is still left to do. Together, we can advance AI governance efficiently and effectively, fostering trust in AI systems at a pace that matches the opportunities ahead.

Footnotes

[1] From Risk to Reward: The Business Case for Responsible AI. clouddamcdnprodep.azureedge.net/gdc/gdcnUixiy/original

[2] The Partnership of the Future - Microsoft's CEO explores how humans and A.I. can work together to solve society's greatest challenges. slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

[3] NIST AI Risk Management Framework. www.nist.gov/itl/ai-risk-management- framework

[4] Microsoft Responsible AI Principles and Approach. www.microsoft.com/en-us/ai/principles-and-approach

[5] Microsoft Responsible AI Standard, v2. blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

[6] Frontier Governance Framework. cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

[7] EU Artificial Intelligence Act. artificialintelligenceact.eu/

[8] Microsoft Enterprise AI Services Code of Conduct. learn.microsoft.com/en-us/legal/ai-code-of-conduct

[9] At Microsoft, Responsible AI CVPs are senior leaders in their division usually reporting directly to a leader at the Executive Vice President level.

[10] AI Ethics and Effects in Engineering and Research committee and working groups leverage top scientific and engineering talent to provide subject-matter expertise on the state-of-the-art and emerging trends regarding the enactment of Microsoft's responsible AI principles.

[11] Microsoft Threat Modeling. www.microsoft.com/en-us/securityengineering/sdl/threatmodeling

[12] Microsoft Learn | Planning red teaming for large language models (LLMs) and their applications. learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

[13] Microsoft Lessons from Red Teaming 100 Generative AI Products. aka.ms/AIRTLessonsPaperf

[14] AI Safety Institute Inspect. inspect.ai-safety-institute.org.uk/

[15] METR Vivaria. vivaria.metr.org/

[16] GitHub Azure/PyRIT. github.com/Azure/PyRIT

[17] Introducing AI Red Teaming Agent: Accelerate your AI safety and security journey with Azure AI Foundry. devblogs.microsoft.com/foundry/ai-red-teaming-agent-preview/

[18] STAC: Sociotechnical Alignment Center. www.microsoft.com/research/group/stac-sociotechnical-alignment-center/

[19] Evaluating Generative AI Systems is a Social Science Measurement Challenge. arxiv.org/pdf/2411.10939

[20] A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. arxiv.org/pdf/2412.01934

[21] Dimensions of Generative AI Evaluation Design. arxiv.org/pdf/2411.12709

[22] EUREKA: Evaluating and Understanding Large Foundation Models. www.microsoft.com/en-us/research/publication/eureka-evaluating-and-understanding-large-foundation-models/

[23] Eureka: Evaluating and understanding progress in AI. www.microsoft.com/en-us/research/blog/eureka-evaluating-and-understanding-progress-in-ai/

[24] Azure AI announces Prompt Shields for Jailbreak and Indirect prompt injection attacks. techcommunity.microsoft.com/blog/azure-ai-services-blog/azure-ai-announces-prompt-shields-for-jailbreak-and-indirect-prompt-injection-at/4099140

[25] Meeting the moment: combating AI deepfakes in elections through today's new tech accord. blogs.microsoft.com/on-the-issues/2024/02/16/ai-deepfakes-elections-munich-tech-accord/

[26] Deceptive AI-Generated Media: Election Misrepresentation. www.microsoft.com/en-us/concern/2024elections

[27] Microsoft on the Issues | MTAC. blogs.microsoft.com/on-the-issues/tag/mtac/

[28] Microsoft and OpenAI launch Societal Resilience Fund. blogs.microsoft.com/on-the-issues/2024/05/07/societal-resilience-fund-open-ai/

[29] MSRC Researcher Portal. msrc.microsoft.com/report/vulnerability/new

[30] How to Report a Compliance Concern | Microsoft Legal. www.microsoft.com/en-us/legal/compliance/sbc/report-a-concern

[31] How AI is changing phishing scams. www.microsoft.com/en-us/microsoft-365-life-hacks/privacy-and-safety/how-ai-changing-phishing-scams

[32] What is PowerShell? learn.microsoft.com/en-us/powershell/scripting/overview?view=powershell-7.5

[33] Mimikatz. www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=HackTool:Win32/Mimikatz

[34] Taking legal action to protect the public from abusive AI-generated content. blogs.microsoft.com/on-the-issues/2025/01/10/taking-legal-action-to-protect-the-public-from-abusive-ai-generated-content/

[35] Disrupting a global cybercrime network abusing generative AI. blogs.microsoft.com/on-the-issues/2025/02/27/disrupting-cybercrime-abusing-gen-ai/

[36] Microsoft Enterprise AI Services Code of Conduct. learn.microsoft.com/en-us/legal/ai-code-of-conduct

[37] Limited Access features for Azure AI services. learn.microsoft.com/en-us/azure/ai-services/cognitive-services-limited-access

[38] Azure AI services. learn.microsoft.com/en-us/legal/cognitive-services/face/transparency-note#government-and-international-organization-use-cases

[39] Welcome to the new Phi-4 models – Microsoft Phi-4-mini & Phi-4-multimodal. techcommunity.microsoft.com/blog/educatordeveloperblog/welcome-to-the-new-phi-4-models---microsoft-phi-4-mini--phi-4-multimodal/4386037

[40] Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. arxiv.org/pdf/2407.13833

[41] Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. arxiv.org/pdf/2407.13833

[42] XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arxiv.org/abs/2308.01263

[43] Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. arxiv.org/pdf/2407.13833

[44] Phi-3 – A Microsoft Collection. huggingface.co/collections/microsoft/phi-3-6626e15e9585a200d2d761e3

[45] Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arxiv.org/pdf/2404.14219

[46] Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. arxiv.org/pdf/2407.13833

[47] Phi-4 - A Microsoft Collection. huggingface.co/collections/microsoft/phi-4-677e9380e514feb5577a40e4

[48] Phi-4 Technical Report. arxiv.org/abs/2412.08905

[49] Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arxiv.org/abs/2503.01743

[50] Microsoft 365 Copilot overview. learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview

[51] Transparency Note for Microsoft 365 Copilot. learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-transparency-note

[52] Overview of optional connected experiences in Office. learn.microsoft.com/en-us/microsoft-365-apps/privacy/optional-connected-experiences

[53] Data, privacy, and security for web search in Microsoft 365 Copilot and Microsoft 365 Copilot Chat. learn.microsoft.com/en-us/copilot/microsoft-365/manage-public-web-access#controls-available-to-manage-web-search

[54] Microsoft 365 Copilot Achieves ISO/IEC 42001:2023 Certification. techcommunity.microsoft.com/blog/microsoft365copilotblog/microsoft-365-copilot-achieves-isoiec-420012023-certification/4397144

[55] Transparency Note for Microsoft Copilot. aka.ms/Copilot-RAI-Transparency

[56] Microsoft Copilot Studio – Generative Orchestrator. learn.microsoft.com/en-us/microsoft-copilot-studio/faqs-generative-orchestration

[57] Microsoft AI for Science Lab. www.microsoft.com/en-us/research/lab/microsoft-research-ai-for-science

[58] AI meets materials discovery - Microsoft Research. www.microsoft.com/en-us/research/story/ai-meets-materials-discovery

[59] Microsoft announces new Copilot Copyright Commitment for customers. blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

[60] Innovate with AI while complying with the EU AI Act. www.microsoft.com/en-us/trust-center/compliance/eu-ai-act

[61] Responsible AI at Microsoft. aka.ms/rai

[62] Azure AI Foundry. azure.microsoft.com/en-us/products/ai-foundry/

[63] Model leaderboards in Azure AI Foundry portal. learn.microsoft.com/en-us/azure/ai-foundry/concepts/model-benchmarks

[64] New GenAI simulation and evaluation tools in Azure AI Studio. techcommunity.microsoft.com/blog/aiplatformblog/new-genai-simulation-and-evaluation-tools-in-azure-ai-studio/4253020

[65] Unlocking the Power of Agentic Applications New Evaluation Metrics for Quality and Safety. devblogs.microsoft.com/foundry/evaluation-metrics-azure-ai-foundry/

[66] Monitor your generative AI applications. learn.microsoft.com/en-us/azure/ai-foundry/how-to/monitor-applications

[67] Correction capability helps revise ungrounded content and hallucinations. techcommunity.microsoft.com/blog/azure-ai-services-blog/correction-capability-helps-revise-ungrounded-content-and-hallucinations/4253281

[68] Multimodal Public Preview Blog. techcommunity.microsoft.com/blog/azure-ai-services-blog/multimodal-public-preview-blog/4253816

[69] Azure Cognitive Services (Azure AI) Application for Embedded Content Safety – Preview. customervoice.microsoft.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR7en2Ais5pxKtso_Pz4b1_xUOFhXTTNZWUNYTlVFOUZMQ1JFSU1QSTFPOCQlQCN0PWcu

[70] EdChat: Pioneering Safe Generative AI in South Australian Classrooms with Azure AI Content Safety. www.microsoft.com/en/customers/story/1751701319789621671-south-australia-department-of-education-azure-ai-content-safety-higher-education-en-australia

[71] How Microsoft and LinkedIn gave users detailed context about media on the professional networking platform. partnershiponai.org/wp-content/uploads/2024/11/case-study-microsoft.pdf

[72] LinkedIn Adopts C2PA Standard. www.linkedin.com/pulse/linkedin-adopts-c2pa-standard-patrick-corrigan-kwldf/?trackingId=4gnjmapwRsmugUNqwj0fRw%3D%3D&src=or-search&veh=www.bing.com%7Cor-search

[73] Introducing greater transparency and control for web search queries in Microsoft 365 Copilot and Microsoft Copilot. techcommunity.microsoft.com/blog/microsoft365copilotblog/introducing-greater-transparency-and-control-for-web-search-queries-in-microsoft/4253080

[74] The European Union Artificial Intelligence Act – The Act Texts. artificialintelligenceact.eu/the-act/

[75] AI Literacy Starting Guide. cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-product-and-services/security/pdf/microsoft-ai-literacy-starting-guide-jan-2025-477720.pdf

[76] AI Skills Navigator. aiskillsnavigator.microsoft.com/en-us

[77] Grant Announcement: Microsoft and Open AI. oats.org/grant-announcemnt-microsoft-and-open-ai/

[78] AI & Society Fellows Program. www.microsoft.com/en-us/research/academic-program/ai-society-fellows/

[79] AI Frontiers – Microsoft Research. www.microsoft.com/en-us/research/lab/ai-frontiers/

[80] Challenges in Human-Agent Communication. Challenges in Human-Agent Communication - Microsoft Research www.microsoft.com/en-us/research/publication/human-agent-interaction-challenges/

[81] Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. www.microsoft.com/en-us/research/publication/supporting-industry-computing-researchers-in-assessing-articulating-and-addressing-the-potential-negative-societal-impact-of-their-work/

[82] The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers.
www.microsoft.com/en-us/research/publication/the-clear-documentation-framework-for-ai-transparency-recommendations-for-practitioners-context-for-policymakers/

[83] From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents.
arxiv.org/pdf/2412.07951

[84] STAC: Sociotechnical Alignment Center - Microsoft Research.
www.microsoft.com/en-us/research/group/stac-sociotechnical-alignment-center/

[85] A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts.
www.microsoft.com/en-us/research/publication/a-shared-standard-for-valid-measurement-of-generative-ai-systems-capabilities-risks-and-impacts/

[86] Dimensions of Generative AI Evaluation Design.
www.microsoft.com/en-us/research/publication/dimensions-of-generative-ai-evaluation-design/

[87] AI Red Teaming Through the Lens of Measurement Theory.
openreview.net/pdf?id=KEggQCeDUA

[88] Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline.
arxiv.org/abs/2501.18493

[89] RedCode: Risky Code Execution and Generation Benchmark for Code Agents.
www.microsoft.com/en-us/research/publication/redcode-multi-dimensional-safety-benchmark-for-code-agents/

[90] Overreliance on AI: Literature Review.
www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[91] A benchmark for evaluating the robustness of LLMs and defenses to indirect prompt injection attacks.
github.com/microsoft/BIPIA

[92] Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models.
www.microsoft.com/en-us/research/publication/benchmarking-and-defending-against-indirect-prompt-injection-attacks-on-large-language-models/

[93] "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust.
www.microsoft.com/en-us/research/publication/im-not-sure-but-examining-the-impact-of-large-language-models-uncertainty-expression-on-user-reliance-and-trust/

[94] Overreliance on AI Literature Review.
aka.ms/genai_reliance

[95] Steering language model refusal with sparse autoencoders.
arxiv.org/pdf/2411.11296

[96] Broadening AI innovation: Microsoft's pledge to the National AI Research Resource pilot.
blogs.microsoft.com/on-the-issues/2024/01/24/national-ai-research-resource-nairr-artificial-intelligence/

[97] Azure Quantum Elements.
quantum.microsoft.com/

[98] Azure AI Foundry: Find the right model to build your custom AI solution.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

[99] Accelerating Foundation Models Research – Projects.
www.microsoft.com/en-us/research/collaboration/accelerating-foundation-models-research/projects/

[100] Microsoft Frontier Governance Framework.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

[101] MLCommons AI Risk and Reliability Working Group.
mlcommons.org/working-groups/ai-risk-reliability/ai-risk-reliability/

[102] Benchmark for general-purpose AI chat model.
ailuminate.mlcommons.org/benchmarks/

[103] Global Governance: Goals and Lessons for AI.
aka.ms/AIGlobalGovernanceBook

[104] Global Perspectives: Responsible AI Fellowship • Stimson Center.
www.stimson.org/project/responsible-ai-fellowship/

[105] Advancing AI Responsibly – Global perspectives on inclusive AI.
unlocked.microsoft.com/responsible-ai/

[106] AI and the SDGs in Latin America and the Caribbean.
scioteca.caf.com/bitstream/handle/123456789/2331/AI%20and%20the%20SDGs%20in%20Latin%20America%20and%20the%20Caribbean.pdf?sequence=1&isAllowed=y

[107] Artificial Dreams.
grandpalais-immersif.fr/en/agenda/evenement/artificial-dreams

[108] National Youth Theatre and Microsoft launch national programme to empower young creatives to use AI.
ukstories.microsoft.com/features/national-youth-theatre-and-microsoft-launch-national-programme-to-empower-young-creatives-to-use-ai/