FRONTIER MODEL FORUM

FY 2024-2025

Annual Report

Table of Contents

Annual Letter from the Executive Director

Since our founding, the FMF has pursued a singular mission: advancing frontier AI safety and security by identifying best practices and standards, supporting scientific progress, and facilitating information-sharing.

I'm proud to share that we've made significant progress over the past year across each of those mandates. By working closely with our member firms and external partners, we've:

- **Advanced best practices.** We've played a key role in pioneering frontier AI frameworks, a new risk management instrument for the most severe risks from the most advanced AI models. Our technical report series is the most comprehensive and in-depth resource to date for the implementation of frontier AI risk management.

- **Supported science and research.** Through our partnership with the AI Safety Fund, we allocated more than $10m in funding for the development of cutting-edge research and methods, including novel safety evaluations such as the Virology Capability Test.

- **Facilitated information-sharing.** We developed a first-of-its-kind information-sharing agreement for vulnerabilities, threats, and concerning capabilities unique to frontier AI, as well as a mechanism for sharing relevant information responsibly and securely.

Thank you to all those who have contributed to each of those initiatives. As grateful as I am for all we've achieved, I'm even more grateful to the many experts and partners – whether in industry, academia, civil society, or the public sector – who have lent their time and insights to our efforts.

We are eager to build on that impact in the year to come. As the pace of AI development continues to accelerate, the importance of working collaboratively across frontier AI developers and sectors will become even more essential. We look forward to continued engagement with all of our stakeholders to ensure that the future of frontier AI remains a safe and secure one.

About FMF

OUR MISSION

The Frontier Model Forum (FMF) is an industry-supported non-profit dedicated to advancing frontier AI safety and security.

The FMF has three core mandates:

1. **Identify best practices** and support standards development for frontier AI safety and security.
2. **Advance the science** of frontier AI safety and security.
3. **Facilitate information sharing** about frontier AI safety and security among government, academia, civil society and industry.

The FMF focuses primarily on managing significant risks to public safety and security, including from chemical, biological, radiological, nuclear (CBRN) and advanced cyber threats. By drawing on the technical and operational expertise of its member firms, the FMF seeks to ensure the most advanced AI systems remain safe and secure so that they can meet society's most pressing needs.

The FMF is committed to working collaboratively with stakeholders across the AI safety and security ecosystem to make AI systems safe, secure, and beneficial.

WORKSTREAMS

The FMF is currently advancing shared understandings of AI safety and security across four workstreams:

AI-Bio

AI Security

AI-Cyber

Frontier AI Frameworks

FMF Members

THE FMF HAS THREE HIGH-LEVEL MEMBERSHIP CRITERIA:

1. Member firms must provide evidence that they have the expertise and resources to develop or deploy at scale frontier AI models.
2. Member firms must have a track record of safety as demonstrated through documentation, safety review processes and research and evaluation.
3. Member firms must have clear intentions to contribute to the FMF.

amazon ANTHROPIC Google Meta Microsoft OpenAI

BOARD MEMBERS	TITLE
Nicole Foster	Director of Amazon Web Services Global AI/ML
Michael Sellitto	Head of Global Affairs, Anthropic
Owen Larter	Senior Director and Head of Frontier Policy and Public Affairs, Google DeepMind
Polina Zvyagina	Director AI Policy and Governance, Meta
Natasha Crampton	Vice President and Chief Responsible AI Officer, Microsoft
Sasha Baker	Head of National Security Policy, OpenAI

Our Workstreams

AI-Bio

AI-Cyber

AI Security

Frontier AI Frameworks

We're committed to developing rigorous and scientifically informed best practices to ensure the safe development and deployment of frontier AI systems.

AI-Bio

This workstream aims to advance the biosafety of leading AI models and systems. Although frontier AI holds significant promise for medicine and the life sciences, it also risks amplifying existing biological threats and introducing novel ones. Understanding how to safely manage those risks is an urgent challenge.

We hosted bi-monthly expert working group convenings with internal and external biosafety experts to identify best practices on threat models, evaluations, thresholds, and safeguards. These convenings helped inform three of our issue briefs on this topic:

- Preliminary Taxonomy of AI-Bio Safety Evaluations (December 2024) that offers an initial taxonomy and definitions for frontier AI safety evaluations specific to the biological domain, categorized across two dimensions: methodology and domain.

- Preliminary Reporting Tiers for AI-Bio Safety Evaluations (March 2025) outlining a three-tiered approach for sharing AI-bio safety evaluation information that aims to attend to both the benefits of greater transparency and the potential risks associated with information and attention hazards. The approach reflects preliminary thinking across FMF member firms about what information from safety evaluations should be shared with the public at large, what should be disclosed within trusted expert networks only, and what should be kept private.

- Frontier AI Biosafety Thresholds (May 2025) that highlights emerging industry consensus on a core biosafety threshold for frontier AI, specifying the point at which an AI model's capabilities in biological domains may require further assessments and/or enhanced safeguards and providing initial guidance for evidence-based threshold determinations.

AI-Cyber and AI Security Workstreams

The AI-Cyber workstream advances the safety and security of leading AI models and systems in the cybersecurity domain. While frontier AI promises significant advances in computing and automation, it also risks amplifying existing cyber threats or introducing novel ones. Understanding and mitigating these risks through robust technical safeguards is vital.

Frontier AI systems also represent high-value targets for adversaries and threat actors. As these systems evolve in capability and complexity, securing the environments in which frontier AI systems are trained and deployed will be an increasingly difficult challenge. Through our AI Security workstream, we aim to meet that challenge by identifying and developing security practices for frontier AI development and deployment.

- We published an issue brief in July 2024 on Foundational Security Practices to provide a baseline understanding of how firms currently approach frontier AI security.
- In an effort to provide a useful starting point for cybersecurity professionals considering how to leverage frontier AI systems for defensive purposes, we also published an issue brief on AI for Cyber Defense in November 2024.

Frontier AI Frameworks Workstream

This workstream aims to advance the development and implementation of robust safety practices for frontier AI models and systems. As AI capabilities continue to advance, comprehensive safety frameworks and evaluations will become increasingly crucial for assessing and addressing potential risks before and during AI development. Understanding how to create, implement, and effectively evaluate safety measures is a critical challenge for the field.

As part of this, we published several issue briefs on the following topics:

- Thresholds for Frontier AI Frameworks
- Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations
- Components of Frontier AI Frameworks
- Early Best Practices for Frontier AI Safety Evaluations

We followed these briefs with a series of technical reports that examine how these frameworks can be implemented effectively across different organizational contexts, including on:

- Frontier Capability Assessments
- Risk Taxonomy and Thresholds for Frontier AI Frameworks
- Frontier Mitigations
- Third-Party Assessments

FRANCE'S AI ACTION SUMMIT - FEBRUARY 2025

We were thrilled to host our side event, "Securing the Future of AI: Harmonizing Frameworks and Approaches," which explored best practices for developing and implementing frontier AI frameworks and explored steps towards global standards alignment. The event began with a discussion of the core components of frontier AI frameworks. Panelists also spoke about the future of these frameworks, including perspectives on open questions regarding their design and implementation. We're very thankful to our fantastic panel of experts: Amanda Craig Deckard, Ashley Zlatinov, Chloe Touzet, Chris Painter, Christian Troncoso, Esther Tetruashvily, Jonas Kgomo, Joslyn Barnhart, Shane Witnov, and Yonadav Shavit.

Advancing the Science

We are fulfilling our core mandate of advancing the science of frontier AI safety and security by supporting the work of the AI Safety Fund (AISF), a collaborative $12 million initiative to accelerate and expand the field of AI safety and security research, and through direct funding to grantees.

The AISF was established in late 2023 as a collaborative initiative among leading AI developers and philanthropic partners. It aims to support and expand the field of AI safety research to promote the responsible development and deployment of frontier models, minimize risks, and enable independent, standardized evaluations of capabilities and safety.

The Meridian Institute initially managed and oversaw the fund. Support for the AISF came from the founding members of the Frontier Model Forum (FMF) – Anthropic, Google, Microsoft, and OpenAI – as well as philanthropic partners such as the Patrick J. McGovern Foundation, the David and Lucile Packard Foundation, Schmidt Sciences, and Jaan Tallinn.

The AISF awarded $4 million in its first round of funding to 12 recipients focused on creating AI biosafety evaluations, developing methodologies to address dangerous capabilities, and identifying deceptive AI behavior. In November 2025, the FMF awarded the second round of AISF grants to 11 recipients, totaling more than $5.2 million, to help address some of the most pressing safety risks associated with frontier AI systems, including biosecurity, cybersecurity, AI agent evaluation, and synthetic content.

After the Meridian Institute announced in June 2025 that it would be winding down its operations, the FMF began managing the fund directly. The remaining funds of the AISF will be used to support narrowly-scoped research projects that target urgent bottlenecks to further progress in AI safety and security.

AISF SUPPORTERS

FMF Founding Member Firms:

- Anthropic
- Google
- Microsoft
- OpenAI

Philanthropic Supporters:

- The Patrick J. McGovern Foundation
- The David and Lucile Packard Foundation
- Schmidt Sciences
- Jaan Tallinn

Advancing the Science *Continued*

AISF GRANTEES

VIROLOGY CAPABILITIES TEST (VCT)

A critical question facing AI developers and regulators is whether frontier models could enable malicious actors to conduct dangerous virology work. To address this, SecureBio's AIxBio team developed the VCT benchmark—an expert-validated assessment tool for evaluating AI models' ability to provide practical troubleshooting guidance for viral research.

This benchmark features 322 multimodal questions covering dual-use research methods, complete with original micrographs and photographs. The dataset is now available to safety teams, regulatory bodies, and academic researchers working on AI-bio safety.

Read Time Magazine article here.

BACTERIAL BIOTHREAT BENCHMARK GENERATION (BBG)

To strengthen biothreat benchmarks for AI model safety assessments, Nemesys Insights developed the Bacterial Biothreat Benchmark Generation (BBG) Framework, a comprehensive assessment system aligned with biosecurity threat chains.

The goal of the framework is to enable the identification of key areas of potential harm along the biothreat chain, allowing for the prioritization of mitigation efforts and enhancing safety more than is possible with more general approaches. The final framework contains approximately 1,000 high-quality bacterial biothreat prompts, complete with implementation tools and risk mitigation recommendations. The benchmark dataset and implementation tool, as well as associated research publications, are forthcoming, pending an information hazard review.

Info-Sharing

The Frontier Model Forum (FMF) was established with a mandate to establish a trusted and secure information-sharing mechanism between industry, civil society, and government.

In March 2025, our member firms signed a first-of-its-kind agreement to facilitate information-sharing about vulnerabilities, threats, and capabilities of concern unique to frontier AI.

AT THIS STAGE, OUR INFORMATION-SHARING COVERS THE FOLLOWING THREE KEY CATEGORIES:

1

Vulnerabilities

Vulnerabilities, weaknesses or exploitable flaws may compromise the safety, security, or intended functionality of frontier AI models. Examples may include jailbreaks, adversarial inputs, data poisoning, or other attempts to bypass model safeguards.

2

Threats

Threats to frontier AI comprise threats directed to the unauthorized access or manipulation of frontier AI models. Examples may include potential threat actors, attack vectors, or cyber-threat indicators.

3

Capabilities of Concern

Capabilities of concern refer to frontier AI capabilities that have the potential to cause large-scale harm to society. Examples may include capabilities related to the development of chemical, biological, radiological, and nuclear (CBRN) threats, offensive cybersecurity attacks, and model autonomy.

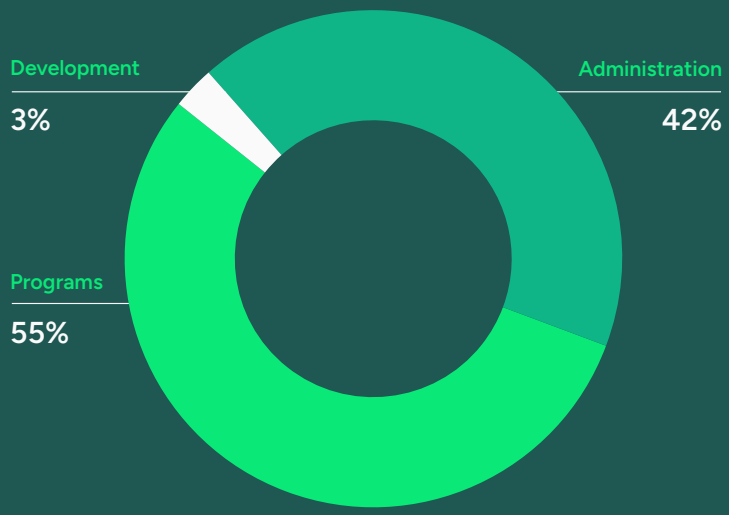
We aim to refine its information-sharing function over time in a way that supports collaboration with the broader frontier AI ecosystem.

Financials

FY2024 OPERATING REVENUE (TOP-LINE FIGURE)

$2,000,000

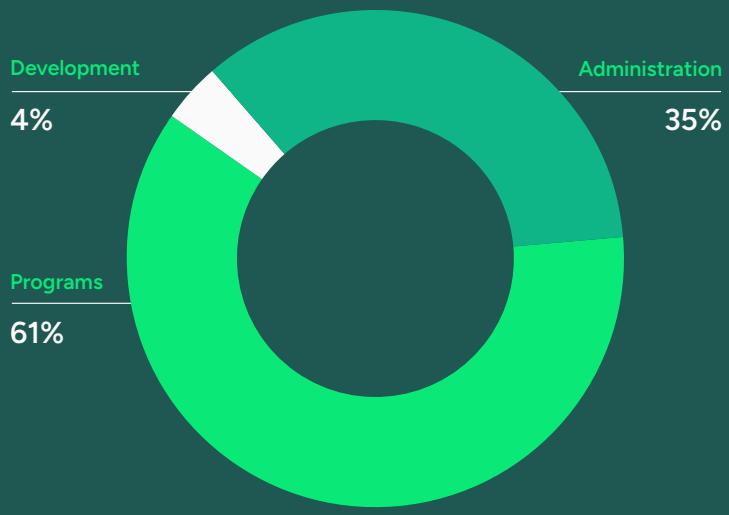
FY2024 EXPENDITURES/EXPENSES



FY2025 OPERATING REVENUE

$3,000,000

FY2025 EXPENDITURES/EXPENSES



FRONTIERMODELFORUM.ORG

Contact Us

INFO@FRONTIERMODELFORUM.ORG