Transparency in Frontier Artificial Intelligence Act (SB 53)

Review of the Frontier Model, Frontier Developer, and Large Frontier Developer Definitions

2027

California Department of Technology
https://cdt.ca.gov

CDT REPORT
Prepared for Members and
Committees of the California State Legislature

Executive Summary

The Transparency in Frontier Artificial Intelligence Act (TFAIA) enacted by Senate Bill 53 (Stats. 2025, ch. 138) took effect on January 1, 2026. It requires developers of the most advanced artificial intelligence (AI) models to publish transparency reports and report critical safety incidents, requires the largest of those developers to publish safety frameworks, and creates whistleblower protections.

What TFAIA requires of CDT

Business and Professions Code (BPC) section 22757.14 directs the California Department of Technology (CDT) to annually assess recent evidence and recommend whether and how to update the statutory definitions of "frontier model," "frontier developer," and "large frontier developer" to ensure they accurately reflect technological developments, scientific literature, and widely accepted national and international standards. In developing its recommendations, CDT must also consider comparable thresholds established under federal law, regulations, or guidance and international standards, including whether alignment with federal definitions would be appropriate and consistent with TFAIA's purposes. CDT must also consider whether the definitions are readily determinable, simple to apply and administer, and externally verifiable.

What CDT Did

CDT reviewed empirical scientific research, published reports, and frontier AI laws, regulations, guidance, and standards in California, other U.S. states, the federal government, and international governments. CDT also ran a public input channel process, to which six organizations submitted input (Appendix A). To ensure inclusive input from stakeholders outlined in statute—including open source, nonprofit, academia, industry, and government communities—CDT held additional consultations with these stakeholders and observed TFAIA-related workshops hosted by Stanford University and UC Berkeley (Appendix B).

Recommendations

Based on its research and stakeholder input, CDT offers the following recommendations on the TFAIA definitions for the Legislature's consideration, as well as additional recommendations to support the implementation and verification of the proposed definitions.

1. **Update the frontier model definition to include model distillation and a capability test.**
2. **Update the large frontier developer definition to include frontier model research and development spending.**
3. **Make no updates to the definition of frontier developer.**
4. **Facilitate collaboration between frontier developers and cloud providers.**
5. **Establish a public registry of large frontier developers and their frontier models.**

Table of Contents

Executive Summary **2**

Table of Contents **3**

Introduction **4**

Section 1: Evolution of the Frontier AI Model Safety, Regulatory, and Evaluation Landscape **5**

California Developments Since Passing of TFAIA 5

International Standards, U.S. State and Federal Law, Guidance, and Regulations 7

Evolution of the Frontier AI Model Risk Profile 15

Technical Evaluations and Benchmarking Processes Used in the Frontier Model Development Lifecycle 16

Emerging Landscape of Third-Party Audits of Frontier AI Models 17

Section 2: Considerations of Definition Complexity **19**

Frontier Model 19

Frontier Developer 21

Large Frontier Developer 21

Section 3: Considerations of Input from Stakeholders **22**

Public Input Channel Comments 22

CDT Stakeholder Workshops 22

University-Hosted Workshops 24

Section 4: Recommendations **26**

Recommendation 1: Update the frontier model definition to include model distillation and a capability test 26

Recommendation 2: Update the large frontier developer definition to include AI R&D spending 27

Recommendation 3: No updates to the definition of frontier developer 27

Recommendation 4: Facilitate collaboration between frontier developers and cloud providers 27

Recommendation 5: Establish a public registry of large frontier developers and their frontier models 27

Appendix A: Input from the Public Input Channel **28**

Appendix B: Stakeholders Consulted **48**

Notes **50**

Introduction

On September 29, 2025, Governor Newsom signed into law Senate Bill 53, the Transparency in Frontier Artificial Intelligence Act (TFAIA). The bill added Chapter 25.1 to the Business and Professions Code (BPC). The chapter relates to the small number of developers building the most advanced, or "frontier," AI models and focuses on the most severe, catastrophic risks those models could pose. Under this chapter:

- Large frontier developers must write, follow, and publish a **frontier AI framework** describing how they assess and manage catastrophic risk.
- Before or when deploying a new or substantially modified frontier model, frontier developers must publish a **transparency report**, and large frontier developers must include summaries of their catastrophic-risk assessments and the extent of any third-party evaluator involvement.
- Frontier developers must report **critical safety incidents** to the Governor's Office of Emergency Services (Cal OES) within 15 days of discovering them, or within 24 hours to an appropriate law enforcement or public safety authority where there is an imminent risk of death or serious physical injury.

TFAIA directs the California Department of Technology (CDT) to assess each year whether the definitions of "frontier model," "frontier developer," and "large frontier developer" continue to reflect developments in frontier AI and to report its findings to the Legislature. These definitions determine which companies and models the law covers.

This report draws on a review of laws, guidance, and technical standards in the United States, the European Union (EU), the United Kingdom, and Asia; written input submitted through CDT's TFAIA public input channel; and stakeholder consultations from open source, non-profit, industry, academia, and government communities.

The report is organized as follows. **Section 1** reviews California's 2026 AI activity and the regulatory, risk, and evaluation landscape. **Section 2** assesses the complexity of each definition. **Section 3** summarizes stakeholder input. **Section 4** sets out CDT's recommendations for the California legislature to consider. Legal citations and technical detail appear in endnotes at the end of the report, and Table 1 defines key terms used throughout the report.

Table 1. Key terms used in this report

Term	Current Statutory Definition
Frontier model	A foundation model trained using more than 10^{26} integer or floating-point operations of computing power. The count includes compute for the original training run and for any later fine-tuning, reinforcement learning, or other material modifications.[1]
Frontier developer	A person or company that has trained, or begun training, a frontier model.
Large frontier developer	A frontier developer that, together with its affiliates, had annual gross revenues of more than $500 million in the preceding calendar year.
Catastrophic risk	A foreseeable and material risk that a frontier model will materially contribute to the death of, or serious injury to, more than 50 people, or more than $1 billion in property damage or loss, arising from a single incident in which the model provides expert-level help with a chemical, biological, radiological, or nuclear weapon; carries out a cyberattack, murder, assault, extortion, or theft without meaningful human oversight; or evades its developer's or user's control.
Critical safety incident	Unauthorized access to, modification of, or exfiltration of a frontier model's weights, or loss of control of a frontier model, that results in death or bodily injury; harm from the materialization of a catastrophic risk; or a frontier model using deceptive techniques to subvert its developer's controls or monitoring, outside an evaluation designed to elicit that behavior, in a way that shows materially increased catastrophic risk.
Compute (FLOP)	The total amount of computation used to train a model, counted in integer or floating-point operations (FLOP). 10^{26} is a 1 followed by 26 zeros.

Section 1: Evolution of the Frontier AI Model Safety, Regulatory, and Evaluation Landscape

California Developments Since Passing of TFAIA

As of this report, the TFAIA remains unamended since its passage in 2025, and the 2026 legislative session built upon this framework.

Building an independent verification system

The Governor signed Senate Bill 813 and Assembly Bill 1405 on September 9, 2026, which direct the Government Operations Agency (GovOps) to establish a process for selecting and regulating independent verification organizations (IVOs) by January 1, 2028, and to develop an AI Auditor Registry by January 1, 2029. Neither law requires frontier developers to obtain an audit or verification.[2]

On September 18, 2026, the Governor issued Executive Order N-9-26[3], accelerating the timelines to May 1, 2027 and December 1, 2027 respectively, and directing GovOps, in consultation with Cal OES and national experts, to submit recommendations by November 16,

2026, on the technical feasibility and potential efficacy of legislative amendments to existing state laws including:

- Placing IVOs onsite at frontier AI laboratories.
- Having IVOs verify developers' TFAIA frameworks, transparency reports, and risk assessments.
- Requiring verified emergency shutdown mechanisms, commonly referred to as a "kill switch" for frontier models.
- Expanding the definition of "critical safety incident" to cover loss-of-control events. Today, loss of control counts only if it causes death or bodily injury.

If adopted, the proposals may amend TFAIA and influence or create related laws in the future.

Other AI Measures Signed in 2026

The Governor also signed other significant AI measures, summarized in Table 2. Executive actions in 2026 included a March order on civil rights, civil liberties, and privacy in state AI procurement[4], a May order on AI's labor impacts, a June Tech Fraud Task Force, and an August AI Cyber Defense Program. These complement the AI cybersecurity playbook that Assembly Bill 979 requires the California Cybersecurity Integration Center to deliver by January 1, 2027.

Table 2. Other AI legislation signed in 2026

Area	Measures signed
Child safety	AB 1709 (addictive features for users under 16); SB 1119, "Adam's Law" (companion chatbot safeguards, with independent child-safety audits from 2029); SB 867 (moratorium on companion-chatbot toys through 2031); AB 2246 (replaces the Age-Appropriate Design Code); AB 1609 (customer-service chatbot disclosure).
Synthetic content	SB 1050 (AI-generated performers in advertising); SB 1000 (removes the one-million-user threshold from the AI Transparency Act while retaining embedded provenance); AB 2713 (platform duties to display and preserve provenance data); SB 1111 (digital replicas).
Workplace	SB 947 (bars sole reliance on automated systems for discipline or termination, effective July 1, 2027); AB 1883 (neural-data and emotion-recognition surveillance); AB 1331 (workplace surveillance); SB 951 (notice when AI causes layoffs).
Health care, biosecurity, and professions	AB 1979 and SB 503 (health care AI); AB 1864 (gene synthesis safeguards); SB 574 (attorneys' use of AI).

International Standards, U.S. State and Federal Law, Guidance, and Regulations

The frontier AI landscape has evolved considerably since the adoption of TFAIA. This section lays out U.S. and international regulations, guidance, and standards on frontier AI since January 2026. Figures 1a and 1b, on the following pages, highlight key laws and regulations, how they interpret key definitions of frontier AI, how they frame risk, and whether they use independent AI safety evaluations.

What the comparison shows

- **U.S. states are converging on California's definitions.** New York amended its Responsible AI Safety and Education (RAISE) Act in March 2026 to track TFAIA's text, and Illinois' SB 315 is also very similar to TFAIA, with the same compute and revenue thresholds.
- **The common gap is verification.** Like California, New York has authority over what developers must disclose but no statutory means to verify it.
- **Jurisdictions are closing that verification gap in different ways.** The EU pairs mandatory reporting with an institution that can demand documents and run its own tests, while Illinois requires large frontier developers to engage an annual independent auditor to verify compliance with the law's framework requirements starting in 2028.
- **Outside of U.S. state laws, risk is often viewed more broadly** at the level of applications, use cases, and impacts to society rather than only the model.

Figure 1a. Binding laws that apply to developers based on model scale

	California (TFAIA)	New York (RAISE Act)	Illinois (AISMA)	European Union (AI Act, GPAI with systemic risk)	South Korea (Korea AI Basic Act)
Status	In force since Jan 2026	Takes effect Jan 2027	Takes effect Jan 2027; framework & audit requirements start in 2028	Applies since Aug 2025; fines since Aug 2026	In force since Jan 2026
Which models are covered	Models above 10^{26} operations; and >$500M in revenue for most obligations	Same as California	Same as California	Models above 10^{25} operations, or designated by the EU	Models trained with at least 10^{26} FLOP that meet other criteria
Developers must have a written safety framework	Yes, for large developers	Yes, for large developers	Yes, for large developers, starting in 2028	Yes, under the Code of Practice	Partial. A risk management system is required.
Developers must report safety incidents to the government	Yes, within 15 days (24 hours if imminent risk)	Yes, within 72 hours (24 hours if imminent risk)	Yes, within 72 hours (24 hours if imminent risk)	Yes, without undue delay (2 to 15 days under the Code of Practice)	Partial. Developers must monitor and respond to incidents.
Government can demand more information, test models, and order recalls	No	No	No	Yes	Partial
Independent review is required	No	No	Partial. Annual compliance audits starting in 2028.	Partial. External evaluations under the Code.	No
Developers must notify government they are covered	No	Yes, for large developers	Yes, for large developers	Yes	No
Maximum penalties	Up to $1M per violation	Up to $1M, or $3M for repeat violations	Up to $1M, or $3M for repeat violations	Up to 3% of worldwide annual revenue or €15M, whichever is higher	Up to KRW 30M (about $20,000)
Annual review on frontier AI thresholds	Yes, CDT annually updates Legislature	No	Partial; from 2029	Partial; as needed	Partial; by decree
Dedicated AI Office	No; four departments / agencies share duties	Yes; DFS office	No; two agencies share duties	Yes; AI Office	No; science ministry

Yes / in progress | Partial or limited | No | Not rated (description only)

Source: CDT analysis of the laws and guidance described in this section. EU ratings include the voluntary General-Purpose AI Code of Practice, which most leading developers have signed. "Partial" means the item is required only in a weaker or incomplete form or is voluntary.

Figure 1b. Jurisdictions without binding safety requirements for frontier developers

	United States (federal)	United Kingdom	China	Singapore
Main instruments	Executive Order 14409 and the White House Accord on Super Intelligence (2026), described as voluntary	No AI statute; voluntary Frontier AI Safety Commitments	Binding rules for generative AI services; non-binding AI Safety Governance Framework; draft cybercrime law and proposed export controls	No AI statute; voluntary IMDA frameworks
Binding rules that reach frontier developers	No frontier-specific rules, but export controls can restrict models	No frontier-specific rules	Security assessment and filing for generative AI services that can shape public opinion; registration before exporting listed AI technologies	No frontier-specific rules
Government testing before release	Pre-release access (up to 30 days), described as voluntary	Voluntary testing by the AI Security Institute	Through the service filing process	None
Voluntary expectations for frontier developers	Under the Accord, internal controls, independent audits, and board oversight	Publish safety frameworks with thresholds for intolerable risk	Draft laws pending	Governance practices that include testing and incident reporting

The United States

New York RAISE Act

New York's RAISE Act is the most similar domestic legislation to TFAIA. The New York Legislature passed the RAISE Act in June 2025, and Governor Hochul signed it on December 19, 2025, with agreed chapter amendments. Those amendments, signed on March 27, 2026, rewrote the law to align with California's text.[5]

As amended, the RAISE Act (in effect January 1, 2027) adopts definitions of "frontier model," "large frontier developer," "catastrophic risk," and "critical safety incident" that are identical to TFAIA's.[6] A large frontier developer must write, implement, comply with, and publish a frontier AI framework, review it at least once per year, and publish a transparency report before, or concurrently with, deploying a new or substantially modified frontier model. The RAISE Act also includes TFAIA's redaction allowance for trade secrets, cybersecurity, public safety, and national security.[7]

New York, like California, has no general power to request technical information from developers, with one structural difference: the RAISE Act creates a dedicated office within the New York Department of Financial Services (DFS) to implement the law. The office establishes

an incident-reporting mechanism, reviews critical safety incident reports, receives confidential quarterly summaries of internal-use catastrophic-risk assessments, maintains a registry of large frontier developers, and can adopt rules.[8] The RAISE Act does not authorize this office to request, require, subpoena, audit, examine, or inspect anything beyond these filings. Nothing in the RAISE Act grants the office the authority to obtain the underlying evaluations, model access, or documentation needed to verify what it is told. Enforcement is by the Attorney General, with civil penalties of up to $1 million for a first violation and $3 million for each subsequent violation.[9]

The RAISE Act does not provide for an annual review of its definitions, as TFAIA does. Measured against TFAIA, the RAISE Act contains the same disclosure regime with a faster clock and a dedicated institution for reporting compliance.[10] Critical safety incidents must be reported within 72 hours, compared with TFAIA's 15 days, with 24-hour notice to an appropriate law enforcement or public safety authority where there is an imminent risk of death or serious injury.[11]

Illinois AI Safety Measures Act

Illinois has enacted the Artificial Intelligence Safety Measures Act (AISMA), effective January 1, 2027, with frontier AI framework and audit obligations beginning January 1, 2028.[12] Like TFAIA, AISMA targets developers of frontier models trained using more than 10^{26} operations, and places most obligations on large frontier developers whose annual gross revenue exceeds $500 million.[13] Unlike TFAIA and the RAISE Act, it requires large frontier developers to retain third-party auditors each year to evaluate compliance with AISMA's frontier AI framework and transparency requirements and to assess internal controls.[14] Redacted audit reports, together with a published high-level summary, must be submitted to the Illinois Attorney General and the Illinois Emergency Management Agency and Office of Homeland Security (IEMA-OHS) within 30 days. Summaries of internal-use catastrophic-risk assessments (every three months) and critical safety incident reports (within 72 hours) also go to the State, and transparency reports are published on the developer's website.[15]

The law includes whistleblower protections and enforcement authorities similar to TFAIA's. Like the RAISE Act, it imposes civil penalties of up to $1 million for a first violation and $3 million for repeat violations.[16] Like New York's law, it does not require annual review of the definitions, although the agency's annual report to the General Assembly and Governor must include any recommended updates to AISMA beginning January 1, 2029.[17] Unlike New York, Illinois has followed California's approach of distributing implementation across existing agencies, namely IEMA-OHS and the Attorney General's Office. On September 22, 2026, Governor Pritzker also created an "AI Cabinet," an advisory group of gubernatorial appointees and senior agency leaders without regulatory or enforcement authority.[18]

Illinois is the only enacted domestic frontier AI law to place a qualified third party between the developer and the state. The auditor must follow generally accepted auditing standards, have demonstrated competence in frontier model safety, and have no financial interest in the developer (nor the developer in the auditor), and payment may not depend on the audit's result. The auditor receives all materials reasonably necessary for the audit, including unredacted versions of published materials, and must state whether the developer "substantially complied" with the requirements of Section 10 of AISMA.[19] Illinois also goes further than TFAIA in two respects: transparency report summaries must be machine-readable "to facilitate verification of model claims," and AISMA establishes a public registry of large frontier developers.[20]

The State's own visibility, however, remains limited. It receives the audit report with the same redactions the developer may apply to any published document, and the audit tests adhere to the developer's own framework.[21] Illinois approaches verification by qualifying and empowering the auditor, rather than by giving the state access to unredacted material.

U.S. Federal Government

This section covers federal measures aimed specifically at frontier developers. It does not cover federal efforts to preempt state AI laws, rules for agencies' own use of AI, or export controls on AI chips. No federal statute specifically regulates frontier AI developers. Executive Order 14409 (June 2, 2026) directs federal agencies to set up a voluntary framework under which developers give the federal government access to "covered frontier models," identified through a classified benchmark of cyber capabilities, for up to 30 days before releasing them to other trusted partners.[22] The federal government has also used binding tools. In June 2026, an export-control directive barring foreign nationals from two Anthropic models led Anthropic to suspend access for all customers for nearly three weeks,[23] and OpenAI said later that month that it was limiting new model releases to a small group of trusted partners at the government's request.[24] The Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) at the National Institute of Standards and Technology (NIST), previously called the Center for AI Standards and Innovation until October 2026, tests frontier models under voluntary agreements with developers.[25]

On September 29, 2026, the White House along with Google, Anthropic, Meta, OpenAI, xAI, and Nvidia signed the White House Accord on Super Intelligence. It says that every company training and deploying frontier models should adopt four layers of controls and audits, consisting of internal monitoring controls, an internal team that checks those controls, an independent external auditor or evaluator, and oversight by an independent board committee.[26] The Accord has no definition of "frontier model," no deadlines, and no enforcement, and it was described at its signing as "morally binding."[27] None of these federal measures requires incident reporting. In October 2026, CAISSI, the National Security Agency

(NSA), and the Cybersecurity and Infrastructure Security Agency (CISA) published key practices for AI incident response that recommend sharing incident information with the government, but the practices are voluntary and set no deadline, while TFAIA requires developers to report critical safety incidents within 15 days of discovering them.[28]

Europe

EU AI Act

The EU Artificial Intelligence Act is a horizontal, risk-tiered AI regulation.[29] The part of the Act most relevant to TFAIA is Chapter V on general-purpose AI (GPAI) models. These obligations have applied since August 2, 2025, and the Commission's enforcement powers, including fines, have applied since August 2, 2026.[30] The Act presumes a GPAI model has "high impact capabilities," and is therefore a model "with systemic risk," when its training compute exceeds 10^{25} FLOP, and the Commission may also designate models by decision.[31] TFAIA sets its frontier model threshold at an order of magnitude higher, 10^{26} operations, and adds a $500 million revenue test for large frontier developers. The two laws cover similar subjects; the AI Act requires model evaluation, risk assessment and mitigation, serious incident reporting, and cybersecurity protections, while TFAIA requires large frontier developers to describe and follow their own approach to these topics in a published framework, and requires all frontier developers to report critical safety incidents.[32]

The largest difference is in empowering the Commission to request technical information from developers. Under the EU AI Act, providers must keep technical documentation on training, testing, and evaluation, and they must provide this information to the EU's AI Office upon request. The Commission may request additional information needed to assess compliance, and when that is insufficient, the AI Office can run its own evaluations of the model, including through independent experts, with access via Application Programming Interfaces (APIs) or source code. The Commission also has the authority to restrict, withdraw, or recall a model, and fining providers up to 3% of worldwide annual turnover or €15 million, whichever is higher, for infringing the GPAI obligations, failing to comply with a request, supplying incorrect or misleading information, or refusing evaluation access.[33]

Unlike the EU AI Act, TFAIA creates no authority for Cal OES, CDT, or the Attorney General to request or compel information beyond the mandated disclosures. In short, the EU pairs mandatory reporting with an institution that has the authority and capacity to verify.

United Kingdom

The United Kingdom has no statute specifically governing generative AI. Its most substantive governance framework comes from a March 2023 white paper, which set out a pro-innovation

approach and five cross-sector principles to be applied by existing regulators under existing law.[34] The white paper stated that "the framework will not involve the creation of a new AI regulator." The government's February 2024 response held to that approach while conceding that current legal frameworks "may not effectively mitigate the risks posed by highly capable general-purpose AI systems."[35] The July 2024 King's Speech pledged legislation for developers of the most powerful AI models, but the May 2026 King's Speech did not call for an AI bill, and as of October 2026 the government had not introduced one.[36] On September 14, 2026, Parliament's Joint Committee on Human Rights called for comprehensive AI legislation and commensurate statutory powers to test and evaluate AI systems before release.

The UK's frontier AI body has no statutory power to request or require technical information from developers. The AI Security Institute (AISI, renamed from the AI Safety Institute in February 2025) describes itself as "a research organization within the UK government's Department for Science, Innovation and Technology" whose work includes testing leading AI systems before public release.

Though the organization has established significant and ongoing relationships with frontier developers, it does not have regulatory authority and its pre-deployment testing is dependent on voluntary agreements. Sectoral regulators maintain the power to collect compulsory information, though even they are limited by their statutes. These regulators only collect catastrophic-risk documentation when a data breach or online safety issue is involved. [37]

As compared to TFAIA, AISI has the technical capacity to test frontier models directly before release, but no legal authority. Access, timing, and scope are all set by one-on-one agreement with the developer. TFAIA provides authority with mandatory frameworks, transparency reports, and incident reports that are backed by penalties. The State receives summaries and self-reports after the fact, rather than pre-release models to test or verify.

Asia

China

The most relevant Chinese guidance on frontier AI risk is the AI Safety Governance Framework 2.0 (September 2025), issued under the guidance of the Cyberspace Administration of China (CAC) by the national technical committee that leads China's AI standards process (TC260).[38] The framework is non-binding but has strong standard-setting effects within China. TC260 released a third edition in September 2026 that adds an annex on managing risks from AI agents.[39] Version 2.0 addresses some of the same risks as TFAIA, including chemical, biological, radiological, and nuclear (CBRN) risks, cyberattacks, and loss of control, but like the EU it takes a more holistic approach. It scores risks by application scenario, level of intelligence, and application scale, and it directs safety assessments at both the model and application levels.

China's binding AI safety rules apply to services rather than to models above a scale threshold. Providers of generative AI services with "public opinion attributes or social mobilization capabilities" must complete a security assessment and file with the CAC, and amendments to the Cybersecurity Law that took effect on January 1, 2026, raised maximum fines to RMB 10 million.[40] Two pending measures address AI cyber-offense capabilities. A draft Cybercrime Prevention and Control Law, published in January 2026 and not yet enacted, would require AI service providers to monitor for, block, and report the use of their services to commit crimes or to generate malicious code in bulk.[41] Export controls proposed on October 16, 2026, would prohibit exporting AI technology for generating malicious code and would restrict exporting AI-assisted vulnerability discovery technology and the core technology, including weights, of large models that meet criteria including training with at least 10^{26} FLOP, the same compute figure TFAIA uses.[42] Exporters of the listed technologies must already register before exporting, and comments on the proposal close on November 15, 2026.[43]

Singapore

Singapore has no legislation directly comparable to TFAIA but has issued relevant non-binding guidance. Compared with California's AI laws, Singapore's guidance focuses more on use cases than on models, emphasizes systematic testing and reliability, draws on safety-engineering principles, and considers a broader range of risks and mitigation strategies than TFAIA covers.

Through the Infocom Media Development Authority (IMDA), the government established AI Verify, a testing and evaluation framework for AI systems developed through a public-private partnership. A notable output is the Model AI Governance Framework for Generative AI (May 2024), which takes a whole-ecosystem approach centered on developer and provider accountability and recommends third-party testing and assurance as well as incident reporting. A similar framework covers agentic AI systems, which TFAIA does not specifically address.[44] IMDA's starter kit for testing applications built on large language models classifies risks by severity (based on domain, affected groups, and systemic factors) and likelihood (based on human oversight and other aspects of sociotechnical design).[45] Thresholds are set in context, based on how systems perform against a suite of tests and scenarios, rather than on a fixed compute or revenue number. This approach is more attentive to the marginal risk posed by an AI system, measured against a baseline of safety that developers determine in context.

South Korea

South Korea has enacted the AI Basic Act, which took effect in January 2026 and is functionally similar to the EU AI Act.[46] However, it has no specific section on general-purpose AI models. Its closest counterpart to TFAIA is Article 32, which requires operators of AI systems trained with at least 10^{26} FLOP that also meet the Enforcement Decree's capability and impact criteria to identify, assess, and mitigate risks, maintain a risk management system, and submit the results

to the Ministry of Science and ICT.[47] The Act separately regulates "high-impact" AI, defined by use in listed sectors such as health care, energy, and hiring, and operators may ask the Ministry to confirm whether a system qualifies.[48] The Ministry may investigate suspected violations of Article 32 and order corrections, and an operator that ignores such an order can be fined up to KRW 30 million. The Ministry has said that for at least the first year it will not impose fines and will investigate only in exceptional cases.[49]

Evolution of the Frontier AI Model Risk Profile

Reorienting AI risk management toward systems

CDT's research and stakeholder input highlight that there are several ways to consider risk as it applies to AI model development and deployment. Since TFAIA was enacted, several approaches to defining catastrophic risk have been under consideration, particularly following a series of 2026 security incidents involving frontier AI developers. In April 2026, unauthorized users reportedly gained access to Anthropic's Claude Mythos Preview, a model withheld from public release because of its cybersecurity capabilities, through a third-party vendor's environment.[50] In July 2026, during an internal cybersecurity evaluation, AI agents driven by OpenAI models bypassed the isolation of their test environment and accessed part of Hugging Face's production infrastructure.[51] Neither incident appears to have met TFAIA's definition of a critical safety incident, yet both illustrate events that concern stakeholders.[52]

Scientific literature and stakeholder input highlighted that, given the pace of AI research, material and societal harm can occur without meeting TFAIA's definition of catastrophic risk. Emerging harms, such as cyber incidents, call for a more systemic orientation to risk that looks at how AI systems are used once deployed, rather than focusing only on models before deployment. Further evidence indicates that understanding the breadth of risk requires considering compute, algorithms, and data bias as they relate to model capabilities, and distinguishing general hazards from the harms that arise through real-world use. Other jurisdictions categorize AI risk in several ways:

- **Tiered risk.** The EU AI Act distinguishes between tiers of risk, with different obligations for providers and deployers: prohibited uses (such as emotion inference in workplaces), high-risk systems, limited-risk systems (subject to transparency duties, such as chatbots and deepfakes)[53], and minimal-risk systems (everything else).[54]
- **Categorical risk.** In binding guidance to federal agencies, the U.S. Office of Management and Budget designates certain AI uses as "high-impact," which triggers minimum risk-management practices such as pre-deployment testing.[55]
- **Material and societal risk.** Comparable global legislation, including the EU AI Act, highlights near-term material and societal risks that TFAIA does not fully cover (listed below).

These material and societal risks include:

- **Behavioral manipulation,** including large-scale inducement of harmful behavior such as self-harm, declining mental health, and distribution of non-consensual imagery.
- **Mass manipulation** that can contribute to violent extremism and acts of terrorism by domestic and international actors, and to events such as mass murder.
- **Information manipulation** that could undermine accurate information on public health, safety, and the administration of elections, or the state's ability to provide services to residents, including through AI-generated video and images.
- **Infringement of individual rights,** such as data privacy in health care, education, the workforce, and law enforcement.

Technical Evaluations and Benchmarking Processes Used in the Frontier Model Development Lifecycle

State-of-the-art tools and practices for assessing frontier models

Assessment of frontier models typically has two components: the frontier developers' own internal safety frameworks, and the evaluation tools those frameworks and third parties use. Major developers—including OpenAI, Anthropic, and Google DeepMind—set capability thresholds in areas such as chemical and biological weapons, cybersecurity, and AI self-improvement, and commit to safety measures as models approach them.[56] The tools that support these frameworks are largely open source.[57] Independent pre-deployment testing is often conducted by various independent third-party evaluators, including those who belong to the AI Evaluation Forum, which evaluates frontier models with OpenAI, Anthropic, Google DeepMind, and Meta, and by CAISSI at NIST, which runs evaluations under voluntary agreements with developers.[58]

Post-deployment monitoring, alerting, and evaluation tools

Post-deployment monitoring is usually conducted internally to developers or driven by media reports. Nothing in the current open-source evaluation tools gives a regulator direct visibility into how frontier models behave in use.[59] One of the primary standardized safety benchmarks is MLCommons AILuminate, built by a working group of more than 100 organizations and covering 12 hazard categories.[60] Harm tracking after the fact is most comprehensively covered by the AI Incident Database, run by the Responsible AI Collaborative from public submissions, and the Organisation for Economic Co-operation and Development (OECD) AI Incidents and Hazards Monitor, which automatically classifies news reports (about 17,600 incidents and hazards).[61]

Compute threshold measurement and estimation

TFAIA sets 10^{26} operations as the compute threshold for frontier models and counts compute from the original training run and any later fine-tuning, reinforcement learning, or other material modifications[62], but no state or federal law, nor any established convention, explains how compute should be estimated. By contrast, the EU AI Act presumes systemic risk above 10^{25} FLOP and the Commission's guidelines uses 10^{23} FLOP as the indicative floor for a general-purpose model. The Act requires notification "without delay and in any event within two weeks" once the systemic-risk line is crossed.[63] The European Commission's Joint Research Centre (JRC) has described two estimation methods, one based on model size and training data and one based on hardware and training time. The JRC recommends counting all computation that shapes the deployed model, raising the EU threshold to 10^{26}, and adjusting thresholds periodically, noting that "training compute thresholds will require periodic adjustments to maintain their objectives."

Epoch AI estimates that more than 30 models from 12 developers had crossed 10^{25} FLOP by June 2025, and that Grok-3 was the first model estimated above 10^{26} (February 2025). In May 2025, Epoch projected roughly 30 models above 10^{26} by 2027 and more than 200 by 2030, concluding that "any regulation based on training compute thresholds would need to cope with a large number of qualifying models, or would need to increase the threshold over time."[64] These projections and estimates are uncertain as AI developers have not published enough information to precisely calculate training FLOP of models.

Cost threshold measurement and estimation

Cost thresholds are also an important consideration. RAISE Act, as passed in June 2025, defined frontier models and large developers partly by compute cost, before the March 2026 chapter amendment replaced those tests with TFAIA's $500 million revenue test.[65] The best-known public methodology for estimating cost is Epoch AI's approach. Using that method, frontier training costs have grown about 2.4 times per year since 2016, and the largest training runs were projected to exceed $1 billion by 2027.[66] Cost-based thresholds are further complicated because cost depends on cloud and/or chips pricing and on whether the developer owns its hardware, and therefore present a certain amount of complexity in verification.

Emerging Landscape of Third-Party Audits of Frontier AI Models

Three regulatory models have introduced a third party that operates independently between a developer and the government. They differ mainly in what information the third party is required, or allowed, to access. As described above, Illinois requires an annual independent

audit of each large frontier developer beginning in 2028. Because the audit covers the developer's own frontier AI framework, it tests adherence to a self-developed standard.

New York's original draft of the RAISE Act also included an annual independent audit, which was dropped before passage. TFAIA requires only that the developer's framework describe the use of third-party evaluators and that transparency reports state "the extent to which third-party evaluators were involved." The EU AI Act's third-party route is narrower: independent "notified bodies" assess only certain high-risk systems, and general-purpose models with systemic risk are not subject to notified-body assessment at all. Their external check is the AI Office's own evaluation capability.[67] Singapore's Model AI Governance Framework for Generative AI states that "third-party testing and assurance often play a complementary role in a trusted ecosystem" and anticipates that an accreditation mechanism could eventually be developed, but its frameworks are voluntary.[68]

The primary challenge for third-party audits is access. Researchers distinguish "black-box" access (querying a model and observing its outputs) from deeper access to a model's internal workings, methodology, code, and data, and conclude that deeper access "allow[s] for substantially more scrutiny than black-box access alone."[69] The UK and U.S. AI Safety Institutes' joint pre-deployment evaluation of OpenAI's o1 model was conducted "during a limited period of pre-deployment access," and the institutes noted that more time and resources could have expanded the scope of their findings.[70] The UK government has identified information access as a structural barrier because firms may "be unwilling to share the required information due to commercial confidentiality concerns," and has suggested "secure enclaves for AI evaluation" as a technical answer.[71]

The lack of clear national and international technical standards for frontier model evaluation is another challenge. Singapore's framework notes that testing requires common benchmarks that do not yet exist, and the UK reports no accredited certification for AI assurance providers. The only published conformity-assessment standard governs bodies that certify AI management systems, meaning organizational processes rather than model capability.[72] Industry often makes the same point against laws such as Illinois': TechNet testified that the state "would effectively be requiring private actors to make highly subjective determinations requiring AI safety compliance without established national standards."[73]

Section 2: Considerations of Definition Complexity

This section highlights the factors that CDT has considered when examining whether and how the definitions should be updated. Each definition is assessed against three tests:[74]

- **Determinability:** Can a developer readily tell whether it is covered, and does the definition capture the right models and companies without including too many or too few?
- **Simplicity:** Is the definition straightforward for developers to apply and for the State to administer?
- **Verifiability:** Can the State of California check whether a developer's determination is correct?

Table 3 summarizes CDT's principal findings for each statutory definition. The discussion that follows provides additional analysis of the considerations underlying those findings.

Table 3. Summary of findings

Definition	Determinability	Simplicity	Verifiability
Frontier model	Can be anticipated before training; but unclear how to count compute, and threshold may miss risk below 10^{26}	Simple proxy, but more accurate capability tests require more evaluation	Developers want guidance on calculating compute; State relies on self-reporting
Frontier developer	Inherits compute-counting ambiguity; attribution unclear for downstream open-weight builders	Simple; follows directly from the frontier model definition	No registry or other means of identifying covered entities
Large frontier developer	Revenue may not reflect risk; well-funded pre-revenue developers are missed, and R&D spending is not well understood	Revenue is simple, but its meaning can be ambiguous	Revenue of private companies is hard to verify independently

Frontier Model

Determinability

CDT's research and stakeholder consultations revealed two challenges for the determinability of TFAIA's frontier model definition: first, ambiguity in how developers should count compute when assessing whether a model is covered; and second, whether the threshold itself captures the models of concern without over- or under-inclusion. Although stakeholders did not agree on how to define "the frontier," there was broad agreement that the existing threshold may not capture all models that could pose catastrophic risk.

One large developer raised concerns that disclosing the exact amount of compute used to train models could reveal proprietary information, and they recommended that the State allow developers to disclose compute in categories, such as "compute bands." Other participants discussed the difficulty of determining exact training compute, particularly when relying on information from third-party cloud compute providers or in-house data centers.

Comparable laws in Illinois and New York use the same 10^{26} threshold. The RAISE Act as originally passed also included a compute-cost test and covered models produced through distillation, but its March 2026 amendment adopted TFAIA's definitions. New York relies on developers' self-reported disclosure statements to maintain its registry of large frontier developers.

In the EU, a compute threshold of 10^{25} FLOP is used as a presumption of systemic risk for general-purpose AI models (the EU's closest equivalent to "frontier" models) rather than as a definition. This allows more flexibility in assessing coverage. Developers can rebut their inclusion, which addresses over-inclusion, and the EU can designate models below the threshold, for example based on independent capability evaluations, which addresses under-inclusion.

Simplicity

Stakeholder input surfaced a tradeoff between the advantages of a simple proxy, such as 10^{26} FLOP, and a more complex approach based on capability evaluations and benchmarking, which would require substantial state expertise and resources to conduct or verify. However, stakeholders suggested that a robust third-party evaluation ecosystem could make more rigorous testing practical for determining whether a model is "frontier."

Verifiability

Stakeholders broadly recommended that the State offer clear guidance on how to calculate compute, since the information a developer needs to support a claim that it is or is not covered is currently unclear. Some stakeholders pointed to the EU, where the EU provides guidance on how post-training work counts toward the total compute used to assess coverage.

In New York, coverage is self-reported, which places the burden of determining coverage on developers and minimizes concerns about over-inclusion. Thresholds are therefore not directly verified, but potential under-inclusion can be addressed through the Attorney General's enforcement powers.

Frontier Developer

CDT's research and analysis of stakeholder input does not suggest that changes to the definition of "frontier developer" are needed in this reporting cycle. The primary focus of this review is the definitions of "frontier model" and "large frontier developer."

Large Frontier Developer

Determinability

Broadly, stakeholder input and CDT's landscape analysis suggested that the current revenue threshold is not a good proxy for risk. While some pointed to the EU AI Act's consideration of scale in determining systemic risk, most stakeholders suggested revisions that would draw a closer connection between risk and the entities developing frontier models.[75]

Both stakeholders and the landscape analysis supported a multi-pronged threshold. Some stakeholders recommended incorporating research and development (compute) costs instead of a revenue threshold. Others pointed out that this approach assumes compute is the best indicator of risk, which remains contested in the scientific literature. One workshop participant pointed to a Michigan bill that would determine coverage using two compute-cost thresholds: one for the training cost of a particular model, and another for compute spending over a 12-month period.[76]

Simplicity

Although a revenue threshold is simple for both regulators and regulated entities to apply, some stakeholders raised concerns about under-inclusion. Some frontier developers, such as startups with billions of dollars in funding focused on research and development, may be well resourced but have little or no revenue. These developers would not be covered under the existing definition, yet may be building highly capable models. Some stakeholders also noted ambiguity in how covered entities should interpret "revenue."

Verifiability

Independently verifying reported revenue is difficult, particularly for private companies. As with the frontier model definition, other jurisdictions rely on self-reporting alongside other verification mechanisms to determine coverage.

Section 3: Considerations of Input from Stakeholders

Public Input Channel Comments

Between March and August 2026, CDT solicited input from academic, open-source, nonprofit, industry, and government stakeholders to complement CDT's research on the TFAIA definitions. This input included public comments submitted through CDT's TFAIA public input channel, listed in Appendix A, as well as additional stakeholder consultations designed to ensure that a diverse and inclusive range of perspectives was represented.

CDT Stakeholder Workshops

In August 2026, CDT conducted a series of stakeholder workshops on the TFAIA definitions. Participants included representatives from industry, academia, nonprofit and public-interest organizations, and the open-source community, listed in Appendix B. Discussions focused on whether the existing definitions continue to reflect developments in frontier AI, how the thresholds can be determined and verified, and whether alternative measures should be considered. Participants weighed administrable definitions against those that more directly reflect model capabilities or catastrophic risk. Views differed on whether the existing compute and revenue thresholds should be retained, revised, or supplemented.

Compute thresholds

Hugging Face questioned whether the existing FLOP threshold may be too arbitrary, while Salesforce noted that the appropriate threshold is likely to change as technology evolves. Reflection AI emphasized model capabilities as another way of thinking about the frontier, stating that "it's more about a capability question than it is a compute one." Black Forest Labs warned against treating the definitions as "one size fits all," noting that different regulatory needs may call for different ways of measuring models. The Center for Democracy & Technology raised the issue that the definition of a “technological frontier” changes over time, noting that models once considered frontier may become dated as newer capabilities emerge.

OpenAI raised concerns that highly granular compute disclosures could reveal commercially sensitive information, suggesting that developers instead report general ranges, or "compute bands." Black Forest Labs and OpenAI both sought clearer guidance on how pre-training, post-training, fine-tuning, and other later development should count toward the threshold. The Center for Democracy & Technology and the Linux Foundation asked how the threshold should apply when developers build on existing open-weight models, with the Center for Democracy &

Technology questioning whether those "building on top of models" would "even have the knowledge to answer questions about what they're building." Downstream developers may not have enough information about the compute used by upstream models to determine if their own work will cross the statutory threshold.

Cost and revenue thresholds

Salesforce questioned if annual gross revenue as a stand-alone measure would capture smaller companies or research organizations with substantial funding and technical capabilities. They noted that significant financial resources may not appear as revenue. OpenAI suggested considering spending on AI development as another measure, and Secure AI Project participants noted that newer companies may develop advanced models before reaching the revenue threshold. On the other hand, Black Forest Labs noted that compute-cost thresholds could be "a lot harder to verify than compute itself" and could require disclosure of additional confidential information, and Google cautioned that research and development spending may not reliably reflect where risk or harm ultimately arises. The Center for Democracy & Technology asked whether coverage should focus primarily on individual models or on the organizations developing them. They suggest greater attention to organizations with significant scale and technical capabilities.

Risk and evaluation

Black Forest Labs questioned whether compute and model size map directly to risk, noting that different categories of risk may require different evaluation approaches and that "a lot of these are system level risks, it's not the model, it's how the model is deployed and distributed." Hugging Face also highlighted the importance of how models are integrated into broader systems when evaluating capabilities and risk. The Center for Democracy & Technology and Tech Equity noted that different categories of risk may require different domain expertise and benchmarks.

Verification and implementation

The Center for Democracy & Technology noted that "developers themselves are likely the only ones who will have the information needed to accurately report their models and compute levels," raising the question of how accurate reporting can be encouraged and independently verified. Tech Equity raised concerns about the consistency and reliability of third-party benchmarks and audits, noting that "a lot of risk measurement is subjective" and that current auditing practices may not yet be mature enough to serve as the primary means of verification. The Linux Foundation emphasized the value of clear, consistent standards, noting that "for good actors who want to comply and are confused about how to, open standards are a good thing." Participants also value coordination across jurisdictions to reduce conflicting standards.

University-Hosted Workshops

In August 2026, Stanford University and UC Berkeley[77] hosted workshops that brought together a similar range of participants to provide input for CDT's report under TFAIA. CDT staff attended as state observers.

Definitions and thresholds

Participants reiterated that compute thresholds provide a predictable and administrable measure yet questioned how well a fixed threshold would continue to reflect model capabilities and catastrophic risk as technology evolves. Attendees noted that "raw compute at training time matters less and less" as development techniques change. Participants discussed whether newer or pre-revenue developers with substantial resources should be covered; if "frontier" should refer only to the most advanced models at a given time; or if differences in model architecture and modality complicate the use of a single threshold; and how pre-training, post-training, reinforcement learning, fine-tuning, and other modifications should affect coverage.

Administrability and verifiability

Administrability and verifiability were described as "two sides of the same coin": regulated entities should be able to determine in advance how the law applies to them, while regulators and the public should be able to evaluate those determinations independently. Participants discussed supplementing compute-based definitions with capability, data, algorithms, and domain-specific criteria to capture lower-compute models that may nevertheless create catastrophic risks. They also asked what information developers already maintain, what additional information may need to be reported, and whether the State has sufficient technical expertise and capacity to evaluate it.

Determinability and its limitations

Participants considered the role of developer attestations, certifications, supporting documentation, and other forms of self-reporting. They noted that developers may not always have strong incentives to describe their own models as more capable, risky, or subject to additional requirements, while some information needed to evaluate a model may be available only to the developer. Participants debated whether documentation or external review could strengthen confidence in coverage decisions while still protecting sensitive or proprietary information.

Pre-deployment testing considerations

Participants noted that internal testing still may not fully predict how a model or system behaves once deployed, and they considered whether post-deployment evidence should play a larger role in understanding model behavior. Experts cautioned against "hyper-focusing on

model capabilities" when assessing real-world harms and emphasized the importance of a broader sociotechnical perspective.

Frontier risk considerations

Participants questioned whether a definition focused primarily on the compute used to train an individual foundation model can account for risks emerging from smaller models, combinations of models, or broader systems. They discussed if and when coverage should remain tied to a fixed threshold or should allow consideration of models below it that exhibit capabilities associated with catastrophic risk, without making coverage overly difficult to determine or administer.

Incident reporting standards and constraints

Participants discussed whether certain model failures, vulnerabilities, or loss-of-control events may provide important safety information even when they do not cause enough physical or financial harm to be reportable. They distinguished conventional cybersecurity vulnerabilities from flaws originating in an AI model itself and stressed that reporting requirements should capture meaningful safety information without becoming too broad to administer.

Third-party evaluators and independent verification organizations

A recurring theme was the expertise and resources required to conduct meaningful evaluations of frontier AI systems, with different categories of risk requiring different subject-matter expertise. Workshop materials noted that current AI evaluations can lack sufficient expertise, consensus, and validity, and called for clearer standards for independent technical evaluation. Participants saw a potential role for academic institutions, independent researchers, open-source communities, government, and third-party evaluators in expanding evaluation capacity, while noting that external evaluators may lack access to information available only to developers.

Section 4: Recommendations

Based on its research and stakeholder input, CDT offers the following recommendations on the TFAIA definitions for the Legislature's consideration, as well as additional recommendations to support the implementation and verification of the proposed definitions.

Recommendation 1: Update the frontier model definition to include model distillation and a capability test

Model distillation

Model distillation is a class of methods for transferring "knowledge" from a frontier model to a smaller student model by training it on outputs of the frontier teacher model.[78] Research has shown that model distillation can produce student models that have comparable performance to their teacher models while reducing the effectiveness of safety measures, including for example, alignment, fine tuning, and guardrails.[79] Therefore, while smaller distilled student models may not themselves meet the compute threshold of 10^26 integer or floating-point operations, they may pose comparable or potentially higher catastrophic risk.

The current definition of a frontier model does not cover distillation. For reference, the RAISE Act, as passed in June 2025 and before its March 2026 chapter amendment, included model distillation as an additional criterion in defining frontier models, covering any model produced by distilling a frontier model at a compute cost of more than $5 million. Based on CDT's research and stakeholder input, distillation should be included in the frontier model definition, given the possibility that a distilled model may exhibit capabilities that can cause catastrophic risk.

Capability test

Stakeholders raised concerns that systems combining several models, none of which meets the threshold on its own, could collectively exhibit capabilities that increase catastrophic risk. Emerging national and international standards and industry practice already use capability thresholds in frontier AI frameworks, where "capabilities" means a system's potential to increase catastrophic risk if deployed without additional safeguards. Capabilities linked to catastrophic risk may arise irrespective of computing power during model training, and should be assessed by experts in the relevant risk areas, determinations which could be verified by independent verification organizations. The definition of frontier model should therefore include a capability test, so that it addresses all models at the frontier of AI development, not only foundation models above the compute threshold.

Recommendation 2: Update the large frontier developer definition to include AI R&D spending

The existing $500 million revenue threshold may not cover developers who are making substantial investments in frontier AI but who are pre-revenue. The definition of large frontier developer should be extended to include a threshold of how much the developer has spent on research and development, including the cost of downstream compute, the cost of compute used to train, evaluate, or serve frontier models before release. 80

Recommendation 3: No updates to the definition of frontier developer

The literature review and stakeholder engagement did not identify any issues with the current statutory definition of frontier developer.

Recommendation 4: Facilitate collaboration between frontier developers and cloud providers

Frontier developers stated that they lack sufficiently detailed information from third-party cloud providers to determine clearly whether they meet the existing thresholds. To improve determinability and verifiability, the legislature or a designated state agency should convene cloud providers and frontier developers to agree on a process for measuring and disclosing frontier model compute, taking into account tenant compute and networking configurations and the definitions in TFAIA. Once complete and approved, this process could be adopted by independent verification organizations for determining and verifying coverage.

Recommendation 5: Establish a public registry of large frontier developers and their frontier models

New York and Illinois both require large frontier developers to file disclosure statements, and Illinois publishes a list of filers, which helps determine which developers are covered. California should also require developers to file a disclosure statement. This filing could list the developers' associated frontier models and could highlight developers' technical documentation relevant to verifying the thresholds that constitute a frontier model. This registry would complement the AI Auditor Registry envisioned under AB 1405, which lists auditors rather than the developers subject to TFAIA.

Appendix A: Input from the Public Input Channel

This appendix presents public input received by CDT through its TFAIA online form. Six organizations provided feedback on the definitions of "frontier model," "frontier developer," and "large frontier developer," along with related considerations for determining coverage under the law. Their responses and supporting evidence are included here to document the varying perspectives submitted for CDT's review and recommendations to the Legislature.

Frontier Model

Question 1

Does the TFAIA definition of "frontier model" accurately reflect the latest technological developments, scientific literature, widely accepted national thresholds, and international standards so that it applies to foundation models at the frontier of artificial intelligence development? *

Response options: Yes / No

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No
Seve Christian Encode AI	No
Thomas Woodside Secure AI Project	No
Bruno Galizzi SaferAI	No
Jimmy Farrell Pour Demain (think tank working on global AI governance)	No
Kavya Pearlman X Reality Safety Intelligence - XRSI	No

Question 2

Please provide your input on the TFAIA definition of "frontier model." *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	While compute thresholds can be a pragmatic starting point, especially while state capacity for more reliable evaluations is still emerging, they suffer from known limitations, and are often set too high to be effective. The current definition of frontier model heavily rests on an arbitrary compute threshold, which while popular, does not have sufficient scientific consensus. This is especially concerning in light of advancements in the field that have proven that powerful models can be built using significantly less compute (see, e.g., , IISS 2025, AlphaEvolve Team 2025, and Quang 2024). New model architectures and training methods can quickly alter the relationship between compute and capability, making compute-based assessment thresholds an unreliable sole proxy. In aggregate, AI capabilities have improved predictably with scale, but it is unclear if resource scaling will improve capabilities at the same rate, which may lead to a shift in attention to other means of improvement. Additionally, compute measures do not offer a direct understanding of a model's general-purpose capability, which is essential for understanding what obligations should be applied and how. Using compute as a proxy overlooks progress in algorithms and the crucial role and impact of training data. Mounting evidence of such innovation in model training, alongside known limitations, nudges us towards setting more conservative and reliable compute thresholds and also calling for the simultaneous consideration of other threshold types for a more accurate assessment of risk. Relying on training compute to gauge a model's capabilities is only suitable when more direct information is unavailable. Regulators assessing a model's general-purpose nature must have access to more accurate, reliable, and repeatable measurement methods that don't have greater scientific consensus. A combination of quantitative and qualitative measures is recommended for evaluating the general-purpose nature of AI models. Consistent task-specific performance metrics along with comprehensive benchmark datasets enable standard statistical analysis and the measurement of factors like accuracy. These metrics should be updated as the understanding of general-purpose models evolves. AI red teaming and other qualitative evaluations offer a more complete view of a model's performance and potential risks. To prevent small yet highly capable general-purpose systems from evading regulations, it is essential to measure the specific qualities that define a general-purpose model.
Seve Christian Encode AI	The definition of “frontier model” provides clarity in statute for which models are covered by the law’s key transparency requirements. While important, this definition does not adequately cover all models with the potential to pose catastrophic risk. We believe that “frontier model” should be defined as both foundation models trained using a quantity of computing power greater than 10^26 FLOP (the current definition) as well as foundation models that are either “reasonably likely to pose catastrophic risks,” or “reasonably likely to possess the ability to automate artificial intelligence research and development,” even if they do not meet the compute threshold. The current definition is not reflective of the original legislative intent or our policy recommendations regarding the importance of an additional qualitative, capability-based approach for scoping models in. As evidenced by the amendments to SB 53 made on July 8, 2025 in Section 22757.12. (a)(1) – allowing for developers, at their discretion, to detail how they exclude models from their safety and security protocols (now frontier AI frameworks) – we find it imperative that developers report when models below a 10^26 FLOP threshold demonstrate risks outlined by this statute. An additional qualitative requirement would help achieve this. We still believe that developers are suited to self-assess which of their foundation models meet this qualitative criteria and that regulatory bodies could help provide guidance if given the statutory ability to do so.

Submitter and organization	Response as submitted
Thomas Woodside Secure AI Project	The current definition of “frontier model” provides a clear compute threshold that determines which models are covered by the law’s key transparency requirements. This is an important starting point, but the current definition does not cover all models with the potential to pose catastrophic risks. Instead, in addition to including foundation models trained using a quantity of computing power greater than 10^26 FLOP (the current definition), it should also include other foundation models that are either “reasonably likely to be capable of posing catastrophic risks,” or “reasonably likely to possess the capability to automate artificial intelligence research and development,” even if they do not meet the compute threshold. Companies could be required to self-assess which of their foundation models meet this qualitative criteria or state officials could be tasked with writing regulations to establish this qualitative threshold in more detail.
Bruno Galizzi SaferAI	We recommend deleting the references to the term ‘foundation’ to categorise the models under scope of the regulation. The term is itself a subject of debate, introducing a layer of uncertainty rather than determining which are the models under scope. For example, Amazon’s definition refers to ‘self-supervised learning to create labels from input data’, which may suggest that foundation models are only those in pre-training. Moreover, the reference to the computational capacity to train the model is already offering a quantitative reference to define the complexity of models under scope. https://aws.amazon.com/what-is/foundation-models/ The reference to 10^26 integer or floating-point operations is in line with the parameters set by the European Commission (10^25), and we recommend keeping monitoring developments to avoid big discrepancies across jurisdictions. We recommend reviewing the potential future implications of using computational capacity as an indication of model capability. In the extensively quoted paper "Algorithmic progress in language models" (Ho, Besiroglu et al., 2024), which looked at 200+ models from 2012–2023, the authors found that the level of compute needed to achieve a given level of performance has halved roughly every 8 months, with a 95% confidence interval of 5 to 14 months. Since the publication of the paper, at least one model has not been publicly released because its cybersecurity capabilities present substantial risk. If we follow this tendency, we may face risks from models with Mythos-level capabilities, requiring substantive less compute than Mythos at training phase. https://arxiv.org/abs/2403.05812 The risk of linking the definition of frontier models mainly to computational capacity, could be increased in a context of model distillation of frontier US models. https://www.iaps.ai/research/ai-distillation-attacks-the-case-for-targeted-government-intervention A recent publication has proposed that model’ memory capacity is replacing FLOPs as a binding hardware constraint developing models. High-bandwidth memory (HBM) has grown from 52% to 63% of total AI chip component spending between Q1 2024 and Q4 2025, and Hyperscalers are already anticipating this in capex guidance. We recommend keeping monitoring this tendency, and eventually integrate it in the definition of frontier models. https://epoch.ai/data-insights/ai-chip-component-cost-shares
Jimmy Farrell Pour Demain (think tank working on global AI governance)	The EU has set a lower threshold at 10^25 FLOP in order to capture a larger amount of models. This is likely to be more future proof especially in the context of algorithmic efficiency is improving, and large-scale risks may come rom increasingly small models. The inclusion of modification and fine-tuning is welcome, however we would have questions about for example: - If transparency obligations are completed on a model exceeding 10^26, and then another applies fine-tuning modifications, does that party only have to report on the fine-tuning themselves? Or the entire model including pre-training? Significant information asymmetries

Submitter and organization	Response as submitted
	exists to make this potentially difficult. It is also essential that this definition is legally robust enough to include internal deployment.
Kavya Pearlman X Reality Safety Intelligence - XRSI	The current TFAIA definition of "frontier model" uses a single training compute threshold of 10^{26} FLOPs as the primary trigger for coverage. This threshold does not accurately reflect the current state of frontier AI development. No publicly known deployed model has been trained at this compute level, meaning the definition as written applies to a class of systems that does not yet exist. At the same time, models trained well below this threshold, including those currently deployed at scale, have demonstrated capabilities that pose catastrophic risks, including autonomous cyberattack execution, zero-day vulnerability discovery, and large-scale influence operations. The definition therefore simultaneously overcaptures future systems that may be benign and undercaptures present systems that are dangerous. A definition grounded in demonstrated capability thresholds rather than training compute alone would more accurately reflect the frontier of AI development as understood by the scientific literature and current international safety evaluation practice.

Question 3

Please provide supportive evidence. *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	Using compute as a proxy for capabilities or risk, on its own, in insufficient and requires pairing with other types of thresholds for a reliable and well rounded estimation (as mentioned in https://cltc.berkeley.edu/publication/intolerable-ai-risk-thresholds/, and https://www.frontiermodelforum.org/issue-briefs/frontier-ai-biosafety-thresholds/). This is particularly critical in light of advancements in the field that have proven that powerful models can be built using significantly less compute (see, https://www.iiss.org/publications/strategic-comments/2025/04/deepseeks-release-of-an-open-weight-frontier-ai-model/). For example, as reported by Google DeepMmind (see, https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/) The inference cost of DeepSseek V3 was estimated to be 12.5 times cheaper than Claude 3.5 Sonnet and more than 15 times cheaper than GPT4o to run. It is also important to point out that methods such as model compression (see, https://medium.com/@nminhquang380/knowledge-distillation-explained-model-compression-49517b039429) can be used to decrease computational requirements while maintaining, and in some cases enhancing, performance. Similarly, a model using a mixture-of-experts (MoE) framework may need less compute to reach the same capabilities as AI using a monolithic model structure (See, https://ai.meta.com/blog/llama-4-multimodal-intelligence/). Additionally, models capable of generating realistic imagery that can be used to create NCII and CSAM often require far less than 10^26 training compute. ADDITION RESOURCES TO CONSIDER: Sara Hooker. On the limitations of compute thresholds as a governance strategy. ArXiv, abs/2407.05694, 2024. URL https://api.semanticscholar.org/ Lennart Heim and Leonie Koessler. Training compute thresholds: Features and functions in ai regulation. (2024). arXiv https://arxiv.org/pdf/2405.10799

Submitter and organization	Response as submitted
	Stephen Casper, Luke Bailey, and Tim Schreier. "Practical principles for AI cost and compute accounting." (2025). arXiv https://arxiv.org/abs/2502.15873 Paul Scharre. Future-Proofing frontier AI regulation: Projecting future compute for frontier AI models. Technical report, CNAS, March 2024. URL https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation Ritwik Gupta, Leah Walker, Rodolfo Corona, Stephanie Fu, Suzanne Petryk, Janet Napolitano, Trevor Darrell, and Andrew W. Reddie. "Data-centric ai governance: Addressing the limitations of model-focused policies." (2024). arXiv https://arxiv.org/abs/2409.17216 Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond et al. "Open problems in technical ai governance." (2024). arXiv https://arxiv.org/abs/2407.14981 Deepika Raman, Nada Madkour, Evan R. Murphy, Krystal Jackson, and Jessica Newman. "Intolerable risk threshold recommendations for artificial intelligence." (2025). arXiv https://arxiv.org/abs/2503.05812
Seve Christian Encode AI	The compute threshold established by SB 53 continues to provide a clear indication for covered developers innovating at the frontier of artificial intelligence and should be retained while the state monitors scientific trends. Most often associated with a 2023 Presidential Executive Order (https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence), this threshold is used now in New York state law and set, pending the Governor's signature, to be referenced in Illinois state law as well: https://www.nbcnews.com/tech/tech-news/illinois-legislature-passes-historic-ai-bill-rcna347191 The European Union's AI Act also utilizes a FLOP threshold as well. Despite the EU compute threshold being one order of magnitude lower than California state law's, however, we do not intend to suggest that a lower threshold is necessary if the aforementioned qualitative threshold is included as well as the statute's current threshold. We also oppose raising the threshold at this time given the scientific uncertainty and lack of publicly available data on compute usage during training of new models known to be demonstrating risk – a very recent and clear example being exemplified by Anthropic's Mythos. (https://www.anthropic.com/glasswing). Establishment of a qualitative threshold in which developers self-assess which models should be accounted for in their policies mitigating catastrophic risk is also a well-documented, common practice among leaders in this space. It is explicitly built into several published safety frameworks of virtually every major frontier AI lab. The core mechanism of these frameworks is typically a tiered, capability threshold system, where companies self-determine which of their models are subject to enhanced safety and security requirements, and which are not. Examples include but are not limited to iterations of: Anthropic's Responsible Scaling Policy and reference to AI Safety Levels (ASL), Open AI's Preparedness Framework and reference to different kinds of capability levels, and Google DeepMind's Frontier Safety Framework and reference to "Critical Capability Levels" (CCLs). To this end, the California Report on Frontier AI Policy released last year – prior to enactment of SB 53 – by Governor Newsom's Joint California Policy Working Group on AI Frontier Models supports the need to supplement a compute threshold with a qualitative threshold. One specific and startling example exemplifying why is found in a 2025 study (https://arxiv.org/abs/2504.16137) cited on page 12 of this report. Here it is cited that OpenAI's o3 model outperformed 94% of virologists. As a result, increasing the capability of

Submitter and organization	Response as submitted
	non-experts to cause biological harm. Not knowing if models like this were trained using at least 10^26 FLOP is evidence that California is in need of a more holistic statutory approach here. A combined compute based and qualitative approach is not a novel concept either in the context of frontier artificial intelligence regulation. The European Union's AI Act's definition of models with "system risk" scopes in models at 10^25 FLOP and *also* models below this threshold that are found (by the companies themselves – as is the case in our definition) to possess the capabilities stipulated in the act.
Thomas Woodside Secure AI Project	The compute threshold of 10^26 should remain a key leg of this definition. This threshold has now become law in New York State [1], and it was previously used in the Biden Administration's AI executive order [2]. This threshold was also used in proposed federal legislation [3] and state legislation currently under consideration in Illinois [4]. While the EU AI Act utilizes a lower threshold of 10^25 FLOP [5], we don't believe it is necessary to lower this threshold if the proposed qualitative threshold is also included, given that many models at this scale do not pose catastrophic risks. We do not believe the evidence supports raising the threshold at this time. For example Anthropic's Claude Mythos model was able to autonomously discover vulnerabilities in major operating systems [6] — a capability highly relevant to SB 53's definition of catastrophic risk — which motivated the company to restrict use of the model to a small group of software companies, rather than release publicly [7]. Mythos might not be classified as a frontier model under a higher compute threshold, as exact compute usage during the training of Mythos is not publicly available. To ensure coverage of models with catastrophic risk potential, this compute threshold must be supplemented with a qualitative threshold. Existing AI systems have already been used by state-sponsored groups to launch autonomous cyber attacks, including in a September 2025 attack executed using Claude Code, as documented by Anthropic [8]. Google [9], Anthropic [10], and OpenAI [11] have each released models that included additional precautions because the companies could not rule out the capability of the models to meaningfully support the development of bioweapons [12]. A 2025 study [13] also found that OpenAI's o3 model outperformed 94% of expert virologists when testing the "capability to troubleshoot complex virology laboratory protocols" and top biosecurity experts said in April 2026 that AI has "meaningfully increased" the risk of biological attacks by "expanding the pool of people who could cause harm" [14] We do not know if these models were trained using at least 10^26 FLOP, and it is essential that these and similar models be covered under the frontier model definition. In addition, we recommend the threshold include a capabilities-based threshold related to the automation of AI research and development. If an AI system is able to conduct research and development towards the end of augmenting its own capabilities with minimal human oversight, it is possible that AI could go through a period of rapid advancement that could contribute to increased catastrophic risks. If AI progress accelerates in this way, new and heightened risks (such as a major jump in cyber capabilities or bioweapons development capabilities) may emerge without warning. Given the potential for such a substantial increase in risk, including models with these capabilities is important for accurately furthering the purposes of the act. OpenAI [15], Anthropic [16], Google [17], and Meta [18] all already reference the automation of AI research and development in their frontier AI frameworks. A definition like this finds precedent in other laws. The EU AI Act also uses a combined qualitative and quantitative threshold in its definition of models with "systemic risk" [19]. The definition presumes that all models trained on greater than 10^25 FLOP are included, while it also covers models below this compute threshold with "high impact capabilities." As in our proposed definition, the EU AI Act also requires companies to assess whether their models meet the qualitative threshold. Additionally, the EU AI Act creates a backstop enabling the European Commission to designate models as meeting the systemic risk

Submitter and organization	Response as submitted
	standard even if companies have not assessed them as such [20] — an option that is not essential but which we would support here, too. The proposed qualitative threshold will be increasingly important in light of recent trends that are rapidly improving the performance of models trained with less compute. First, model performance can improve based on the amount of compute employed at the time of use, which is known as inference compute. With the use of more inference compute [21], models can think through problems in more detail before providing an output, resulting in higher performance [22] and potential increases in risk for any given size of model. Second, companies are increasingly capable of developing high-performance models using less training compute through distillation [23], a process that involves training these models directly on the outputs of models trained with more compute. Third, companies are achieving higher model performance for a given level of training compute through algorithmic advancements [24]. While the training compute threshold remains an important starting point, a qualitative threshold will provide a backstop ensuring that models that pose a catastrophic risk are covered, even if they are trained with less compute. [1] https://www.nysenate.gov/legislation/laws/GBS/A44-B [2] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence [3] https://www.congress.gov/bill/119th-congress/senate-bill/2938/text [4] https://www.ilga.gov/Legislation/BillStatus/FullText?LegDocId=210927&DocName=10400SB0315sam001&DocNum=315&DocTypeID=SB&LegID=157797&GAID=18&SessionID=114&SpecSess=&Session= [5] https://artificialintelligenceact.eu/high-level-summary/ [6] https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf [7] https://www.anthropic.com/glasswing [8] https://www.anthropic.com/news/disrupting-AI-espionage?ref=zef.plus [9] https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf [10] https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf [11] https://cdn.openai.com/gpt-5-system-card.pdf [12] https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf [13] https://arxiv.org/abs/2504.16137 [14] https://www.nytimes.com/2026/04/29/us/ai-chatbots-biological-weapons.html [15] https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf [16] https://trust.anthropic.com/resources?s=gi5v45ke7aezh7b04e82s&name=anthropic-frontier-compliance-framework-[feb-2026-] [17] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf [18] https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2/ [19] https://artificialintelligenceact.eu/article/51/ [20] https://artificialintelligenceact.eu/article/52/ [21] https://openai.com/index/learning-to-reason-with-llms/ [22] https://www.tobyord.com/writing/mostly-inference-scaling [23] https://www.gov.uk/government/publications/ai-insights/ai-insights-model-distillation-html [24] https://epoch.ai/blog/algorithmic-progress-in-language-models
Bruno Galizzi SaferAI	Shared above

Submitter and organization	Response as submitted
Jimmy Farrell Pour Demain (think tank working on global AI governance)	EU rules on thresholds: https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers Increasing evidence of harmful fine-tuning using low levels of compute: https://arxiv.org/abs/2604.24902 On internal deployment in the EU context: https://www.apolloresearch.ai/governance/internal-deployment-eu-ai-act/
Kavya Pearlman X Reality Safety Intelligence - XRSI	Supporting evidence: Current frontier models including GPT-4o, Claude 3.5 Sonnet, and Gemini Ultra are estimated to have been trained at approximately 10^{24} to 10^{25} FLOPs, one to two orders of magnitude below the statutory threshold. (Epoch AI, Training Compute Trends, 2025) Anthropic's Claude Mythos Preview, evaluated by the UK AI Security Institute in April 2026, demonstrated autonomous offensive cyber capabilities including zero-day discovery and end-to-end cyberattack execution at levels the developer itself assessed as too dangerous to deploy, at compute levels below 10^{26} FLOPs. (AISI, Evaluation of Claude Mythos Preview, April 2026) The UK AISI and METR have both published evaluation frameworks demonstrating that dangerous capability thresholds can be operationalized through benchmark-based assessments independent of training compute. (AISI, Early Lessons from Evaluating Frontier AI Systems, October 2024) Epoch AI projected that open-weight models would surpass 10^{26} FLOPs of training compute by late 2025, suggesting the threshold will become relevant but that the intervening regulatory vacuum is consequential. (Epoch AI, January 2025)

Frontier Developer

Question 4

Does the TFAIA definition of "frontier developer" accurately reflect the latest technological developments, scientific literature, widely accepted national thresholds, and international standards so that it applies to developers of frontier models who are themselves at the frontier of artificial intelligence development? *

Response options: Yes / No

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No
Seve Christian Encode AI	No
Thomas Woodside Secure AI Project	No

Submitter and organization	Response as submitted
Bruno Galizzi SaferAI	No
Jimmy Farrell Pour Demain (think tank working on global AI governance)	Yes
Kavya Pearlman X Reality Safety Intelligence - XRSI	Yes

Question 5

Please provide your input on the TFAIA definition of "frontier developer." *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	We do not agree with the need for the distinction between a "frontier developer" and a "large frontier developer" because the potential risks associated with both these entities are rather identical. We recommend that the bill be refocused solely on large frontier developers as defined below.
Seve Christian Encode AI	The concept of a "frontier developer" should be eliminated, with the statute refocused solely on large frontier developers as defined below.
Thomas Woodside Secure AI Project	The concept of a "frontier developer" should be eliminated, with the statute refocused solely on large frontier developers as defined below.
Bruno Galizzi SaferAI	We recommend clarifying the extension of the term 'trained'. It is unclear if the current definition includes organisations that are only fine-tuning models or not. There may be organisations that have not been involved in the initial 'pre-training' of the model, but they execute heavy reinforcement learning for specific models. It is unclear if these developers would be in scope of the regulation. Similarly, we recommend clarifying the phrase 'initiating the training of'. It is unclear if it is referring to an organisation that kicked off a training programme of a model and it is going through it, or if the expression refers to an organisation that has commissioned a model externally. As a point of comparison, in the EU AI Act, any organisation that 'has a GPAI model developed' is in scope, including organisations that commission others to develop a model before placing it in the market themselves. We recommend considering this approach as a way to clarify interpretation. We recommend clarifying the notion of 'intended use' of specific models, providing an operationalisable definition as "intended use" is notoriously difficult to document or define in concrete terms. Article 3 in the EU AI Act, could offer some guidance defining 'intended purpose' as 'use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation'.
Jimmy Farrell Pour Demain (think tank	N/A

Submitter and organization	Response as submitted
working on global AI governance)	
Kavya Pearlman X Reality Safety Intelligence - XRSI	The current definition of "frontier developer" is derivative of the frontier model definition and inherits its primary weakness: it is compute-bound rather than capability-bound. A developer who produces a model with demonstrated catastrophic risk potential but trained below the 10^{26} FLOP threshold is not a frontier developer under TFAIA, regardless of the actual risk their system poses. Additionally, the definition does not account for developers who build dangerous capabilities through fine-tuning, agentic scaffolding, or model distillation on top of base models trained by others. A company that takes a 10^{24} FLOP base model and fine-tunes it into a system with autonomous offensive cyber capabilities is not captured by the current definition, even though the resulting system may be more dangerous than a natively trained frontier model.

Question 6

Please provide supporting evidence. *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	Given the lack of consensus on singular thresholds to determine the 'frontier' nature of a model, we recommend that it not be the requisite criteria to qualify a frontier developer. We provide alternative measures under the feedback on the next question.
Seve Christian Encode AI	SB 53 is intended to be targeted at AI companies at the frontier of AI research. If, as we propose below, the revenue threshold in the definition of "large frontier developer" is replaced with a computational spend threshold, then "large frontier developer" would capture all companies that are developing the most advanced frontier models. There would be no need for a separate definition of "frontier developer" and the statute should be amended as such.
Thomas Woodside Secure AI Project	SB 53 is intended to be targeted at AI companies at the frontier of AI research. If, as we propose below, the revenue threshold in the definition of "large frontier developer" is replaced with a computational spend threshold, then "large frontier developer" would capture all companies that are developing the most advanced frontier models. There would be no need for a separate definition of "frontier developer" and the statute should be amended as such.
Bruno Galizzi SaferAI	Shared above
Jimmy Farrell Pour Demain (think tank working on global AI governance)	N/A
Kavya Pearlman X Reality Safety Intelligence - XRSI	Supporting evidence: The Frontier Model Forum's April 2026 report on managing advanced cyber risks in frontier AI frameworks explicitly identifies fine-tuning and agentic scaffolding as pathways through which dangerous capabilities can emerge in systems below conventional compute

Submitter and organization	Response as submitted
	thresholds. (Frontier Model Forum, April 2026) NIST's AI Risk Management Framework identifies capability evaluation as the appropriate basis for risk classification, not training compute alone. (NIST AI RMF 1.0, 2023) The EU AI Act defines high-risk AI systems by use case and capability, not by training compute, providing a complementary definitional model. (EU AI Act, Annex III, 2024)

Large Frontier Developer

Question 7

Does the TFAIA definition of "large frontier developer" accurately reflect the latest technological developments, scientific literature, widely accepted national thresholds, and international standards so that it applies to well-resourced frontier developers? *

Response options: Yes / No

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No
Seve Christian Encode AI	No
Thomas Woodside Secure AI Project	No
Bruno Galizzi SaferAI	No
Jimmy Farrell Pour Demain (think tank working on global AI governance)	Yes
Kavya Pearlman X Reality Safety Intelligence - XRSI	Yes

Question 8

Please provide your input on the TFAIA definition of "large frontier developer." *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	The current definition of "large frontier developer" relies primarily on company revenue as a proxy for organizational scale and capability. While revenue may be a useful indicator of commercial success, it is not necessarily a reliable measure of frontier AI development activity. Additionally, anchoring the definition to revenue may create loopholes for developers building models for non-commercial (e.g., research) purposes. We therefore support a compute-based definition of large frontier developers, which would capture both commercial and research-focused organizations developing highly capable models, including those that may deploy systems internally before generating significant revenue. This approach better aligns regulatory obligations with the actual sources of frontier AI risk.
Seve Christian Encode AI	The current definition of large frontier developer established by SB 53, defined as companies making over $500M in annual revenue, fails to account for whether this revenue is actually linked to frontier AI development and fails to cover companies that may instead focus on research and development instead of making their models commercially available to the public. In the latter example, these companies may not meet the revenue threshold while still working specifically to develop incredibly advanced models with unique risk profiles. A solution to this issue would be to amend the definition of of large frontier developer to be specific to developers who, in a calendar year, “utilized a quantity of computing power to develop foundation models the cost of which would exceed five hundred million dollars ($500,000,000) if measured using the prevailing market price of cloud computing in the United States at the time the computing power was used.”
Thomas Woodside Secure AI Project	The existing definition of large frontier developer is predicated on company revenue, with the goal of identifying the most well-resourced companies at the frontier of AI development. But this definition does not cover well-resourced companies that may focus on research and development rather than commercial deployment. These companies may not meet the revenue threshold, but they may be developing models that pose significant catastrophic risk. Instead, large frontier developer should be defined as a developer that “in the last twelve calendar months, used a quantity of computing power to develop frontier models the cost of which would exceed five hundred million dollars ($500,000,000) if measured using the prevailing market price of cloud computing in the United States at the time the computing power was used.” Alternatively, this gap could be addressed by updating the definition to cover companies that meet either this proposed compute spend threshold or the existing revenue threshold.
Bruno Galizzi SaferAI	We identify a potentially problematic discrepancy between the rationale used to define ‘frontier developers’, linked to the definition of frontier models and therefore to computational capacities, and the definition of ‘large frontier developers’, which refers to annual gross revenues. It is unclear how these two parameters are linked to coherently describe the frontier AI development market. We recommend relaxing the gross revenue threshold used in the present definition to characterise a ‘large frontier developer’. Even if market leaders such as OpenAI or Anthropic, are clearly beyond this 500 million US dollars bar, reporting revenues at around 20 billion (see references below), we presume that an increasing number of companies will have the capacity to develop models with top performing capabilities without reaching this threshold. https://www.bloomberg.com/news/articles/2026-03-03/anthropic-nears-20-billion-revenue-run-rate-amid-pentagon-feud?embedded-checkout=true https://finance.yahoo.com/news/openai-cfo-says-annualized-revenue-173519097.html?guccounter=1 Finally, we recommend clarifying the definition of the term ‘affiliates’, as it is not clear if it refers to ‘shareholders’, ‘stakeholders’, ‘collaborators’, or other types of associations.

Submitter and organization	Response as submitted
Jimmy Farrell Pour Demain (think tank working on global AI governance)	Agree that using a revenue floor is useful way to avoid unnecessary burden on small players that are significantly less likely to produce large-scale risks. However, given that more risks are likely to arise from smaller open-source and potentially distilled models, this should be flexible to be revised to include smaller players when necessary.
Kavya Pearlman X Reality Safety Intelligence - XRSI	Yes But Partially. The $500M annual revenue threshold for "large frontier developer" is a pragmatic and defensible design choice that appropriately narrows the law's compliance burden to a small number of well-resourced actors. This is one of TFAIA's genuine strengths compared to its predecessor SB 1047. However, the revenue threshold functions as a static proxy for resource capacity rather than a dynamic measure of risk. A developer with $499M in revenue building a system with autonomous CBRN uplift capability would not be a large frontier developer under the current definition. As the AI industry matures and capable systems become cheaper to develop, the revenue threshold will increasingly fail to capture well-resourced actors who are genuinely at the frontier. A recommended modification would be to define "large frontier developer" through either the revenue threshold OR demonstrated capability thresholds on evaluated risk benchmarks, whichever is triggered first. This would ensure that the compliance burden remains appropriately targeted while closing the gap created by declining development costs.

Question 9

Please provide supporting evidence. *

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	Compute expenditure is closely linked to the development of frontier models and has increasingly been proposed as a practical threshold for identifying entities that warrant additional governance requirements. Advanced models are often tested and used internally before public release and may operate under different oversight conditions than externally deployed systems. As a result, a framework that focuses on development capacity rather than commercial revenue is better aligned with the objective of identifying developers capable of creating frontier systems that could generate severe harms. In this way, a compute-based threshold would create a more reliable regulatory trigger that better reflects the developers' resource capacity. We support the submission from the Secure AI Project team that tracks the need for a "cloud computing benchmark" to best operationalize this effort.
Seve Christian Encode AI	Company revenue does not accurately account for investments that are specific to artificial intelligence. A threshold based on compute spend would be more accurate to assess which companies are developing models at the frontier of capability. This is also not a novel regulatory concept to be proposed. Last year, while Senate Bill 53 was moving through the legislature, a report (https://carnegieendowment.org/research/2025/07/artificial-intelligence-regulation-united-states) was published by the Carnegie Endowment for International Peace by a former AI Advisor at the White House and a Yale Law Professor that came to the same conclusion. And similarly, Anthropic proposed (https://www.anthropic.com/news/the-need-for-transparency-in-frontier-ai) a transparency framework that specifically utilized research and

Submitter and organization	Response as submitted
	development spend as a metric to scope in companies that should be subject to relevant transparency requirements. Current trends in artificial intelligence scaling laws demonstrate that the more data that a model is trained on, utilizing larger amounts of computational power and resources, the more capable it becomes (Sastry et al. https://arxiv.org/abs/2402.08797). Recognizing that some at the cutting edge of this innovation do not release commercial products and would not be scoped into the requirements of state law under SB 53, the need to address this is urgent. The most prime example illustrating the need to modify this definition in this way is Safe Super Intelligence Inc. (https://ssi.inc/). Founded by former OpenAI cofounder Ilya Sutskever, the sole mission of this project is to create a model with superintelligence: a currently hypothetical agent that can outperform human intelligence in almost every regard and that many in the field of AI research question can indeed be pursued safely (Bengio et al. https://arxiv.org/abs/2502.15657) Safe Superintelligence, despite not receiving revenue, could develop highly capable models that are deployed internally and present risk. Yet, they would not meet requirements of the current definition of a larger frontier developer. California State law now recognizes that the internal use of models can still present risk and developers are now required by statute (BPC § 22757.13. (a)(4)) to report critical safety incidents that are associated with internal use and should make sure that this requirement comprehensively addresses the issue at hand.
Thomas Woodside Secure AI Project	Because company revenue does not indicate AI-specific investment, it is a poor measure of a company's status as a large frontier developer. Instead, a threshold based compute spend (as defined above) will more accurately capture the desired qualities. A 2025 report by Dean Ball, a former AI advisor at the White House, and Ketan Ramakrishnan, a Professor at Yale Law School, also identified compute spend as a promising threshold for determining which entities should be subject to certain regulatory requirements [1]. This change is necessary to capture companies that do not release or prioritize commercial products until reaching higher capability levels (like former OpenAI co-founder Ilya Sutskever's Safe Superintelligence Inc. [2]). These companies could develop highly capable models and deploy them internally without receiving any revenue. This is a pressing gap to address, as internal use may pose risks [3] that are contemplated directly in SB 53 22757.12(a)(4), 22757.12(a)(10), and 22757.12(d) as well as the California Report on Frontier AI Policy [4]. These risks are compounded by the fact that internal models often have capabilities that exceed the publicly-available frontier, and they are often used without the safeguards that companies employ for externally deployed models. A May 2025 report from AI evaluation nonprofit METR found that developers frequently give AI agents unrestricted permissions and allow them to take actions without human approval during internal use, modes that are activated with the commands "dangerously-skip-permissions" or "yolo" [5]. Under the current definition, companies without a focus on commercialization may not be a large frontier developer, even if they bring in major investments (which can exceed $1 billion [6]) and spend significant sums on training compute. With this in mind, we suggest a $500 million compute training threshold. This sum exceeds the amounts spent on the largest training runs prior to 2025 [7], but likely covers the compute spend needed to develop the most advanced models currently available in 2026, in line with the intention for SB 53's requirements to cover the next generation of frontier models to be released after the bill's passage. The updated definition we suggest would ensure that companies are required to have frontier AI frameworks if they develop models that pose a catastrophic risk. We suggest using a cloud computing benchmark for compute spend for several reasons. First, many AI companies purchase the hardware needed for computation for use over

Submitter and organization	Response as submitted
	multiple years, requiring complex amortization. Second, AI companies sometimes receive compute as part of complex equity deals that are complicated to value [8]. Finally, it is possible that using actual compute costs could encourage offshoring datacenters to locations like the UAE, which would harm American competitiveness. By using a cloud computing baseline, all companies would be placed on a level playing field regardless of their particular cost structures. The legislature could also consider authorizing an agency to promulgate guidance or regulations for the calculation of prevailing cloud computing prices. [1] https://carnegieendowment.org/research/2025/07/artificial-intelligence-regulation-united-states [2] https://ssi.inc/ [3] https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/ [4] https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025--The-California-Report-on-Frontier-AI-Policy.pdf [5] https://metr.org/risk-report-feb-mar-2026.pdf#page=14 [6] https://www.reuters.com/technology/artificial-intelligence/openai-co-founder-sutskevers-new-safety-focused-ai-startup-ssi-raises-1-billion-2024-09-04/ [7] https://epoch.ai/trends#training-runs [8] https://www.aboutamazon.com/news/company-news/amazon-invests-additional-5-billion-anthropic-ai
Bruno Galizzi SaferAI	Shared above
Jimmy Farrell Pour Demain (think tank working on global AI governance)	N/A
Kavya Pearlman X Reality Safety Intelligence - XRSI	Supporting evidence: The cost of training frontier-capable models has declined significantly due to architectural efficiency gains. Anthropic's Constitutional AI and Google DeepMind's efficiency research both demonstrate that dangerous capability levels are achievable at substantially lower compute and cost than the current thresholds assume. The EU AI Act's "general purpose AI model with systemic risk" designation is triggered by training compute above 10^{25} FLOPs, a lower threshold than TFAIA, and includes provisions for the European AI Office to designate additional models based on capability assessment regardless of compute. (EU AI Act, Article 51, 2024)

Optional input

Question 10

Please provide input on thresholds used in international standards for management of catastrophic risk.

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No response provided.
Seve Christian Encode AI	No response provided.
Thomas Woodside Secure AI Project	Few international standards address AI catastrophic risk. The International Standards Organization has published a standard (ISO/IEC 42001:2023 [1]) focused on the responsible use and development of AI, but this standard does not address catastrophic risks. While not a true international standard, the European Union's AI Act creates a standard for EU nations and applies to all companies that offer relevant models in the EU market [2]. Since most major AI companies seek to offer their models in the EU, the AI Act has global influence. As described above, the EU AI Act creates a similar model definition of "general-purpose AI model with systemic risk" [3], which is analogous to SB 53's "frontier model" delineation. Like SB 53, the EU AI Act begins with a compute threshold but also incorporates a qualitative definition targeting high risk models that fall under this threshold. This approach aligns with our proposed update to the "frontier model" definition. [1] https://www.iso.org/standard/42001 [2] https://artificialintelligenceact.eu/high-level-summary/ [3] https://artificialintelligenceact.eu/article/51/
Bruno Galizzi SaferAI	No response provided.
Jimmy Farrell Pour Demain (think tank working on global AI governance)	The safety and security chapter of the EU Code of Practice outlines how providers should define risk thresholds and risk acceptability criteria. Whilst this is a good step, it leaves the decision up to providers which should be defined by the rules themselves: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai The TFAIA should go beyond this by defining clear thresholds, perhaps taking inspiration from Anthropic's AI Safety Levels, but making them a public, enforceable standard: https://www.anthropic.com/news/anthropics-responsible-scaling-policy
Kavya Pearlman X Reality Safety Intelligence - XRSI	International standards thresholds for catastrophic risk: The EU AI Act uses 10^{25} FLOPs as the trigger for systemic risk designation, one order of magnitude lower than TFAIA. The UK AI Security Institute's evaluation framework uses demonstrated capability benchmarks across four domains: offensive cyber capability, CBRN uplift, autonomous deception, and critical infrastructure attack potential. The OECD AI Principles define high-risk AI by impact domain and capability, not compute. California's threshold is the highest of any major jurisdiction and is therefore the least protective relative to any current international standard.

Question 11

Please provide input on thresholds used in federal law, guidance, or regulations for the management of catastrophic risk.

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No response provided.
Seve Christian Encode AI	No response provided.
Thomas Woodside Secure AI Project	There are currently no binding federal obligations to address catastrophic risk from AI models. Many leading AI developers have voluntarily agreed to provide model access to the Center for AI Standards and Innovation (CAISI), housed within the National Institute for Standards and Technology (NIST) at the U.S. Department of Commerce, for evaluations focused on national security risks [1]. However, if there is a threshold above which models are provided for testing, it has not been publicly reported. [1] https://www.cnbc.com/2026/05/05/ai-oversight-trump-google-microsoft-xai.html
Bruno Galizzi SaferAI	No response provided.
Jimmy Farrell Pour Demain (think tank working on global AI governance)	No response provided.
Kavya Pearlman X Reality Safety Intelligence - XRSI	Federal law and guidance thresholds: The Biden Executive Order on AI (October 2023) used 10^{26} FLOPs as its notification threshold for dual-use foundation models, which is the source of TFAIA's threshold. However, that order also included provisions for the Secretary of Commerce to update thresholds based on developments in interpretability and evaluations, an adaptive mechanism TFAIA does not currently include. NIST SP 800-218A and the forthcoming NIST SP 800-218 R2 on Secure Software Development Framework are both moving toward capability-based risk classification for AI-integrated systems.

Question 12

Please provide input on any recommended changes to the law, if necessary, to enable a person to determine as clearly and early as possible, before beginning to train or deploy a foundation model, whether that person will be subject to the definition as a frontier developer or as a large frontier developer.

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No response provided.
Seve Christian Encode AI	No response provided.
Thomas Woodside Secure AI Project	We agree that it is valuable for developers to be able to determine whether they are covered by the relevant definitions prior to training or deploying a model. Since compute

Submitter and organization	Response as submitted
	spend on AI training is more directly within a developer's control (and more narrowly defined) than total revenue, we think that our proposed update to the "large frontier developer" definition will make this threshold more predictable for developers. Our proposal to update the frontier model definition by including an additional qualitative threshold could increase uncertainty for some developers who have spent above the cost threshold but whose models are below the compute threshold. While this is a downside, we believe that most developers will be clearly below or above this threshold, so this potential uncertainty will apply to few developers. We also believe that it is worth adding one additional layer of potential uncertainty to limit the chance that models below the compute threshold pose catastrophic risks and remain uncovered by key transparency requirements.
Bruno Galizzi SaferAI	No response provided.
Jimmy Farrell Pour Demain (think tank working on global AI governance)	Perhaps there needs to be allowances for punitive action if the pre-training FLOP ends up going over 10^26 (or the recommended 10^25), when the provider didn't plan for that to be the case and therefore didn't notify authorities.
Kavya Pearlman X Reality Safety Intelligence - XRSI	Recommended changes for early determination of coverage: Add a transitional coverage tier at 10^{23} FLOPs with a reduced but binding obligation set covering current top-tier commercial models. This would close the regulatory vacuum for the next several years while the industry approaches the 10^{26} threshold. Pair this with a capability-based override: any model that scores above a defined threshold on AISI-aligned evaluation benchmarks for offensive cyber capability or CBRN uplift is covered regardless of training compute.

Question 13

Please provide input on any recommended changes to the law, if necessary, to allow simpler determinations of whether a person or foundation model is covered.

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No response provided.
Seve Christian Encode AI	No response provided.
Thomas Woodside Secure AI Project	While simplicity is an important goal, we do not believe that additional changes need to be made to increase the simplicity of the existing definitions. Our proposal to shift the large frontier developer definition to a compute spend threshold is comparable in simplicity to the current, revenue-based threshold. Our proposed frontier model definition does trade off some simplicity by adding in the qualitative threshold. However, we believe that this proposed threshold is the simplest way to include models under the compute threshold that pose catastrophic risks. Furthermore, we contend that the benefits of ensuring coverage of risky models will exceed the costs of one additional layer of complexity in this definition.

Submitter and organization	Response as submitted
Bruno Galizzi SaferAI	No response provided.
Jimmy Farrell Pour Demain (think tank working on global AI governance)	No response provided.
Kavya Pearlman X Reality Safety Intelligence - XRSI	Recommended changes for simpler coverage determinations: Publish a standing list of covered models maintained by CDT, updated quarterly, based on both compute estimates and capability evaluation results. This shifts the verification burden from individual developers self-reporting to a state-maintained registry, similar to how the FDA maintains drug approval databases. Developers could petition for inclusion or exclusion with supporting evaluation evidence.

Question 14

Please provide input on any recommended changes to the law, if necessary, to ensure that whether a person or foundation model is covered is verifiable by parties other than the frontier developer.

Submitter and organization	Response as submitted
Deepika Raman UC Berkeley Center for Long Term Cybersecurity (CLTC)	No response provided.
Seve Christian Encode AI	No response provided.
Thomas Woodside Secure AI Project	At this time, we are not aware of a workable definition of frontier model that would be easily verifiable by third parties that have access only to publicly available information. Statistics such as the quantity of compute used to train AI models are not publicly available, and alternative metrics such as AI benchmark results are often dependent on specific testing scenarios, and therefore can be difficult to verify externally. Even revenue is not reported publicly for many large AI companies that are privately held. In light of this, if the legislature wishes to ensure verifiability of thresholds, it could consider requiring a third party auditor to verify that, similarly to what Illinois has considered in SB 315. [1] https://ilga.gov/Legislation/BillStatus/FullText?LegDocId=197587&DocName=10400SB0315eng&DocNum=315&DocTypeID=SB&LegID=157797&GAID=18&SessionID=114&SpecSess=&Session=
Bruno Galizzi SaferAI	No response provided.
Jimmy Farrell Pour Demain (think tank	No response provided.

Submitter and organization	Response as submitted
working on global AI governance)	
Kavya Pearlman X Reality Safety Intelligence - XRSI	Recommended changes to ensure coverage is verifiable by third parties: Require covered developers to submit compute attestations certified by independent third-party auditors at the time of model registration, using a standardized reporting template developed by CDT in consultation with NIST. Mandate that pre-deployment capability evaluations be conducted by accredited independent evaluators from a CDT-maintained approved list, with results submitted directly to CDT rather than routed through the developer. This removes the self-certification problem that currently makes coverage determinations effectively unverifiable.

Appendix B: Stakeholders Consulted

The following tables identify the organizations and individual contacts included in CDT's outreach and consultations on the TFAIA . The academic, nonprofit and open-source, and industry lists reflect those who participated in our consultations. The invitations included a broader list of stakeholders. Government organizations consulted are listed separately.

Academia

Organization	Individual contact(s)
University of California, Berkeley	Deirdre K. Mulligan
University of Pennsylvania	Danaé Metaxa
University of Michigan	Abigail Z. Jacobs
UC Berkeley Center for Long Term Cybersecurity (CLTC)	Deepika Raman

Nonprofits and open source

Organization	Individual contact(s)
Center for AI and Digital Policy	Merve Hickok Christabel Randolph
Secure AI Project	Thomas Woodside
METR	Matt Putz Jasmine Dhaliwal
Center for Democracy & Technology	Kevin Bankston
TechEquity	Vincent Le
The Linux Foundation	Steve Winslow
Encode AI	Seve Christian
SaferAI	Bruno Galizzi
Pour Demain	Jimmy Farrell
X Reality Safety Intelligence (XRSI)	Kavya Pearlman

Industry

Organization	Individual contact(s)
OpenAI	Traci Lee Meng Jia Yang
Google	Elizabeth Banker Alice Friend Elli Sakhaee
Amazon	Kevin Snell
SpaceX / xAI	Jordan Rodell
Salesforce	Rachel Gillum
Hugging Face	Ian Reynolds
Black Forest Labs	Ben Brooks
Reflection AI	Joseph Brandifino
Proteus Strategies	Derek Slater

Government organizations consulted

Organization
National Institute of Standards and Technology (NIST)
European Commission
EU AI Office
UK Government
State of New York

Notes

[1] Bus. & Prof. Code § 22757.11(i).

[2] SB 813 (Stats. 2026, ch. 179), adding Gov. Code §§ 8898–8898.4; AB 1405 (Stats. 2026, ch. 178), adding Gov. Code §§ 11549.80–11549.86. The Governor signed both bills on September 9, 2026. SB 813 provides that it does not require anyone to engage an IVO or undergo a covered AI audit (Gov. Code § 8898.4(a)(3)).

[3] Cal. Exec. Order N-9-26 (Sept. 18, 2026). Paragraphs 1 and 2 of the order set May 1, 2027, and December 1, 2027, as the deadlines for GovOps to complete the requirements of Gov. Code §§ 8898.1 and 11549.82(a), and paragraph 3 asks for recommendations by November 16, 2026.

[4] Cal. Exec. Order N-5-26 (Mar. 30, 2026).

[5] S6953B/A6453B (2025); chapter amendment S8828 (2026).

[6] N.Y. Gen. Bus. Law art. 44-B, §§ 1420–1429. A "frontier model" is "a foundation model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations" (§ 1420(9)); a "large frontier developer" is one that "together with its affiliates collectively had annual gross revenues in excess of five hundred million dollars" (§ 1420(10)); "catastrophic risk" is a foreseeable and material risk of death or serious injury to more than fifty people or more than $1 billion in property damage (§ 1420(3)); and "critical safety incident" covers unauthorized access to model weights, materialization of a catastrophic risk, loss of control, and deceptive model behavior (§ 1420(4)).

[7] §§ 1421(1), 1421(2)(A), 1421(3), 1421(5).

[8] §§ 1420(16), 1422(1), 1422(2)(A), 1422(4), 1428, 1429. Failure to file a disclosure statement carries a penalty of $1,000 per day. The office may also designate substantially equivalent federal requirements as satisfying the reporting section (§ 1422(8)–(9)). The Act does not address whether DFS's general powers under the Financial Services Law apply to a frontier developer that is not otherwise a regulated entity.

[9] § 1427(1).

[10] The reporting mechanism captures the same four elements as TFAIA's (§ 1422(1) mirrors Bus. & Prof. Code § 22757.13(a)); incident and internal-use reports are exempt from New York's Freedom of Information Law, as California's are from the Public Records Act (§ 1422(6)); and the office must publish an annual anonymized, aggregated report beginning January 1, 2028, a year after Cal OES's first (§ 1422(7)).

[11] § 1422(3).

[12] SB 315, P.A. 104-0538.

[13] § 5.

[14] § 10(d).

[15] §§ 10(c), 10(d)(4)(A), 10(e), 15.

[16] §§ 20, 25.

[17] § 15(g).

[18] Ill. Exec. Order No. 2026-07 (Sept. 22, 2026).

[19] § 10(d)(1)–(2). The developer may impose security protocols such as on-premises review and restrictions on note-taking. The report must also describe material deviations, assess internal controls, list personnel and conflict-of-interest procedures, set out the methodology, and carry the lead auditor's certifying signature.

[20] §§ 10(c)(3), 18. From January 1, 2027, no large frontier developer may develop, deploy, or operate a frontier model in Illinois without a current disclosure statement, renewed annually and backed by a $1,000-per-day penalty. The Agency publishes the list of filers and collects fees to administer the Act.

[21] §§ 10(d)(4)(B), 10(g).

[22] Exec. Order No. 14409, Promoting Advanced Artificial Intelligence Innovation and Security, § 3 (June 2, 2026), 91 Fed. Reg. 34565 (June 5, 2026).

[23] Anthropic, "Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5" (June 12, 2026), https://www.anthropic.com/news/fable-mythos-access; Ashley Capoot, "Anthropic Says Trump Admin Has Lifted Export Controls on Claude Fable 5 and Mythos 5," CNBC (June 30, 2026). According to CNBC, Anthropic announced on June 30, 2026, that the Department of Commerce had lifted the controls, made Fable 5 available to all users again on July 1, and had restored Mythos 5 access for some U.S. organizations after government approval on June 26.

[24] Ashley Capoot, "Anthropic Says Trump Admin Has Lifted Export Controls on Claude Fable 5 and Mythos 5," CNBC (June 30, 2026), reporting OpenAI's announcement of June 26, 2026.

[25] NIST, Center for Advancing Innovation and Standards for Super Intelligence (CAISSI), https://www.nist.gov/caissi. NIST adopted the new name following Exec. Order No. 14434, Inaugurating the Era of Super Intelligence (Sept. 29, 2026), 91 Fed. Reg. 63129 (Oct. 2, 2026).

[26] White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities (Sept. 29, 2026), posted by President Trump on Truth Social; text reproduced by The American Presidency Project, https://www.presidency.ucsb.edu/documents/white-house-accord-super-intelligence.

[27] Joe Walsh, "Trump and Major AI Executives Sign 'Morally Binding' Voluntary Controls," CBS News (Sept. 29, 2026), https://www.cbsnews.com/news/trump-ai-constitution-tech-execs-openai-anthropic-voluntary-controls/.

[28] NIST CAISSI, NSA Super Intelligence Security Center, and CISA, "Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation," PP-26-4396 (Oct. 2026), practices 9.1–9.3, https://www.nist.gov/document/keypracticesforincidentreadinessandresponseinsuperintelligence-trainingandevaluation. The document describes the practices as voluntary and states that it is not intended as advice for compliance, regulatory, or legal purposes.

[29] Regulation (EU) 2024/1689, as amended by the Digital Omnibus on AI, Regulation (EU) 2026/1744 (in force July 27, 2026).

[30] Models placed on the market before August 2, 2025, have until August 2, 2027, to comply. Because the Omnibus deferred obligations for high-risk AI systems to December 2, 2027, and August 2, 2028 (Art. 113, as amended), the GPAI chapter is the operative comparison.

[31] Art. 51(1)(b), 51(2).

[32] Compare Art. 55(1) (state-of-the-art model evaluation, including documented adversarial testing; assessment and mitigation of systemic risks; serious incident reporting; and cybersecurity) with Bus. & Prof. Code §§ 22757.12 (frontier AI frameworks and transparency reports) and 22757.13 (critical safety incident reporting).

[33] Arts. 53(1)(a), 53(3), 91–93, 101(1). For high-risk systems, market surveillance authorities also have "full access" to documentation and to training, validation, and testing datasets, and to source code on reasoned request (Art. 74(12)–(13)). The Omnibus gave the AI Office equivalent powers over AI systems built on a provider's own GPAI model (new Arts. 75(1), 75a).

[34] The principles are safety, security, and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress. Department for Science, Innovation and Technology (DSIT), *A pro-innovation approach to AI regulation* (Mar. 29, 2023).

[35] Government response, CP 1019 (Feb. 2024), which also stated "We will not 'rush to regulate'" and set out "a pro-innovation case for further targeted binding requirements on the small number of organisations developing highly capable general-purpose AI systems."

[36] Patricia Figueiredo, "UK AI Laws and Security Institute Powers Remain Undecided, Govt Official Says," MLex (Oct. 14, 2026), reporting a government official's testimony to a parliamentary committee on October 13, 2026.

[37] The Information Commissioner's Office may issue information notices under data protection law (Data Protection Act 2018, s. 142); the Competition and Markets Authority may require information "relevant to a digital markets function" (Digital Markets, Competition and Consumers Act 2024, s. 69); and Ofcom may issue

information requests under the Online Safety Act 2023, backed by fines of up to £18 million or 10% of qualifying worldwide revenue. Ofcom's November 2024 open letter confirmed that generative AI services fall within the Online Safety Act where they operate as user-to-user, search, or pornography services.

[38] Cyberspace Administration of China, release notice for the AI Safety Governance Framework 2.0, issued by TC260 (Sept. 15, 2025). The analysis in this report reflects version 2.0.

[39] National Technical Committee 260 on Cybersecurity (TC260), AI Safety Governance Framework 3.0 (Sept. 14, 2026), released at the opening of China's 2026 National Cybersecurity Week.

[40] Cyberspace Administration of China et al., Interim Measures for the Management of Generative Artificial Intelligence Services, art. 17 (effective Aug. 15, 2023); NPC Standing Committee, Decision on Amending the Cybersecurity Law (adopted Oct. 28, 2025; effective Jan. 1, 2026).

[41] Ministry of Public Security, Cybercrime Prevention and Control Law (Draft for Public Comment), art. 42 (Jan. 31, 2026). The comment period closed on March 2, 2026.

[42] Ministry of Commerce and Ministry of Science and Technology, Announcement No. 48 of 2026 (Oct. 16, 2026); Ministry of Commerce, spokesperson's answers on the revision of the Catalogue of Technologies Prohibited or Restricted from Export (Oct. 16, 2026). See Announcement No. 48, Attachment 1 (draft list of technologies), items 266402J (prohibited; AI-driven malicious code generation and mutation), 266404X (restricted; AI-assisted vulnerability discovery), and 266504X (restricted; large models).

[43] Announcement No. 48 of 2026, parts 1–3; Ministry of Commerce, spokesperson's answers (Oct. 16, 2026).

[44] IMDA, Model AI Governance Framework for Agentic AI (Jan. 22, 2026).

[45] IMDA, Starter Kit for Testing LLM-based Applications for Safety and Reliability (Jan. 2026).

[46] Framework Act on the Development of Artificial Intelligence and Establishment of a Foundation for Trustworthiness (Act No. 20676, promulgated Jan. 21, 2025; in force Jan. 22, 2026). The Enforcement Decree adopted on January 21, 2026, sets a 10^{26} FLOP threshold for the Article 32 safety obligations. An amending Act that took effect on July 21, 2026, added provisions on public procurement of AI products, support for AI use by vulnerable groups, and support for AI startups. Lee Myeonghwan, "Revised AI Basic Act Takes Effect Today," Asia Business Daily (July 21, 2026).

[47] AI Basic Act, art. 32; Enforcement Decree of the AI Basic Act, art. 24 (adopted Jan. 21, 2026; in force Jan. 22, 2026).

[48] AI Basic Act, arts. 2(4) and 33–35.

[49] AI Basic Act, arts. 40 and 43. On the grace period, see Shin & Kim, "AI Basic Act Update: Enforcement and Key Implications" (Feb. 11, 2026), summarizing the Ministry of Science and ICT's announcement that for at least one year it will not impose fines and will conduct fact-finding investigations only in exceptional cases, such as those involving loss of life or human rights violations.

[50] Bloomberg, "Anthropic's Mythos Model Is Being Accessed by Unauthorized Users" (Apr. 21, 2026). Anthropic stated that it found no evidence its own systems were affected.

[51] Hugging Face, "Security incident disclosure: July 2026" (July 16, 2026); OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation" (July 21, 2026). Both companies reported the incident as contained, and Hugging Face found no tampering with its public models or datasets.

[52] The Mythos Preview incident involved access by unauthorized people and caused no reported injury. The definition also covers a frontier model that uses deceptive techniques to subvert its developer's controls, even without injury, but OpenAI's quarterly internal-use report states that the models involved in the Hugging Face incident fall outside that report's scope, which covers OpenAI's frontier models. Bus. & Prof. Code §§ 22757.11(d), 22757.12(d); OpenAI, "Periodic Summary of Internal Use Catastrophic-Risk Assessments" (June 4 to September 4, 2026), section 2.1.

[53] Regulation (EU) 2024/1689, Art. 50.

[54] Regulation (EU) 2024/1689, Art. 5 (prohibited practices); Chapter III and Annex III (high-risk systems).

[55] OMB Memorandum M-25-21, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust* (2025).

[56] OpenAI, Preparedness Framework v2 (Apr. 15, 2025), assesses biological and chemical, cybersecurity, and AI self-improvement capabilities against "High" and "Critical" thresholds, backed by expert red-teaming and third-party "deep dives." Anthropic, Responsible Scaling Policy v3.4 (July 8, 2026), ties its ASL-3 standards to CBRN and AI research and development thresholds, with capability evaluations at six-month intervals. Google DeepMind, Frontier Safety Framework 3.1 (Sept. 22, 2025, updated Apr. 2026), sets Critical Capability Levels across CBRN, cyber, machine-learning research acceleration, and harmful manipulation. The Frontier Model Forum (Amazon, Anthropic, Google, Meta, Microsoft, and OpenAI) publishes shared procedures for "Frontier Capability Assessments" in the same domains.

[57] These include the UK AI Security Institute's Inspect framework (more than 200 pre-built evaluations); EleutherAI's lm-evaluation-harness (more than 60 benchmarks; the back end of Hugging Face's Open LLM Leaderboard); and Stanford CRFM's HELM.

[58] METR publishes a "time horizon" metric based on the length of tasks models can complete. CAISSI's 2026 evaluations included DeepSeek V4 Pro, GLM-5.2, and Kimi K3.

[59] Developer tools include Anthropic's Clio (Dec. 2024), "an automated analysis tool that enables privacy-preserving analysis of real-world language model use," and OpenAI's threat-intelligence reports on disrupted malicious use, published since February 2024. Open-source red-teaming tools include NVIDIA's garak and Microsoft's PyRIT.

[60] AILuminate v1.0 includes 59,624 text prompts and results for 109 models.

[61] *Trusted frontier model metric repositories.* As of September 2026, only a few repositories have earned broad industry, government, and academic trust. These include Epoch AI's Notable AI Models database and Benchmarking Hub, and Stanford's annual AI Index Report. Epoch AI's database covers more than 3,600 models with training-compute and cost estimates, each labeled by confidence (within 3x, 10x, or 30x). Its Benchmarking Hub independently re-runs GPQA Diamond, FrontierMath, SWE-bench Verified, and other benchmarks, and tracks 80 external benchmarks. HELM provides standardized multi-metric scores, and Arena (formerly LMArena) ranks models through crowdsourced comparisons. Government evaluators (CAISSI and AISI) and METR publish aggregated reports rather than publicly searchable data.

[62] Bus. & Prof. Code § 22757.11(i).

[63] Regulation (EU) 2024/1689, Arts. 51(2) and 52(1); European Commission, Guidelines on the Scope of the Obligations for Providers of General-Purpose AI Models (July 18, 2025).

[64] Epoch AI, "Over 30 AI Models Have Been Trained at the Scale of GPT-4" (data insight, as of June 2025); Epoch AI, "How Many AI Models Will Exceed Compute Thresholds?" (May 30, 2025), which identifies Grok-3 as the first model in Epoch's dataset estimated above 10^{26} FLOP and gives the projections and quotation in the text.

[65] As passed, the Act covered models trained using 10^{26} operations "the compute cost of which exceeds one hundred million dollars," with a $5 million test for distilled models, and defined a large developer by "over one hundred million dollars in compute costs in aggregate."

[66] Epoch AI, "How much does it cost to train frontier AI models?" (June 3, 2024). The method combines amortized hardware depreciation and energy over the training run, cross-checked against cloud rental prices. Hardware accounts for 47–67% of total development cost, research staff for 29–49%, and energy for 2–6%.

[67] Under Art. 43(1), only Annex III point 1 (biometric) systems go to a notified body, and only where harmonized standards or common specifications were not applied; systems in Annex III points 2 to 8 use internal control (Art. 43(2)). Notified bodies must be independent of the provider and maintain confidentiality (Arts. 31, 34, 78).

[68] IMDA and AI Verify Foundation, Model AI Governance Framework for Generative AI (May 2024). The separate Model AI Governance Framework for Agentic AI (Jan. 22, 2026) adds baseline testing and human checkpoints.

[69] Casper et al., "Black-Box Access is Insufficient for Rigorous AI Audits," ACM FAccT (2024).

[70] U.S. AI Safety Institute and UK AI Safety Institute, *Pre-Deployment Evaluation of OpenAI's o1 Model* (Dec. 2024).

[71] DSIT, *Trusted Third-Party AI Assurance Roadmap* (Sept. 2025). Secure enclaves have been piloted by the UK AI Security Institute, Anthropic, and OpenMined.

[72] ISO/IEC 42006:2025 (July 2025), which applies to certification under ISO/IEC 42001.

[73] TechNet, testimony on Illinois SB 315 (2026).

[74]These three criteria should be interpreted as a starting place for defining complexity, and not a definitive list. As noted on page 38 of the California Report on Frontier AI Policy (June 17, 2025), "a number of additional considerations influence threshold design," and the report highlights three: determination time, measurability, and external verifiability. https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf

[75] Regulation (EU) 2024/1689, Annex XIII.

[76] Michigan HB 4668 (2025), which remains pending in committee as of October 2026.

[77] Stanford Institute for Human-Centered Artificial Intelligence & UC Berkeley, *Implementing the Transparency in Frontier Artificial Intelligence Act* (2026), https://hai.stanford.edu/assets/files/hai-berkeley-implementing-the-tfaia-report.pdf.

[78] Hinton et. al. 2015

[79] Kim & Rush 2016, Gu et al. 2024, Agarwal et al, 2024. Xu et al. 2025