Future of Life Institute

AI Safety Index

Winter 2025

December 2025

Available online at: futureoflife.org/index

Contact us: policy@futureoflife.org

future of life INSTITUTE

Contents

About the Organization: The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Learn more at futureoflife.org.

1 Executive Summary

The Future of Life Institute's AI Safety Index provides an independent assessment of eight leading AI companies' efforts to manage both immediate harms and catastrophic risks from advanced AI systems. Conducted with an expert review panel of distinguished AI researchers and governance specialists, this third evaluation reveals an industry struggling to keep pace with its own rapid capability advances—with critical gaps in risk management and safety planning that threaten our ability to control increasingly powerful AI systems.

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Overall Grade	C+	C+	C	D	D	D	D	D-
Score	2.67	2.31	2.08	1.17	1.12	1.10	1.02	0.98
Risk Assessment (6 indicators)	B	B	C+	D	D+	D	D	D
Current Harms (7 indicators)	C+	C-	C	F	D	D+	D+	D+
Safety Frameworks (4 indicators)	C+	C+	C+	D+	D-	D+	F	F
Existential Safety (4 indicators)	D	D	D	F	F	F	F	F
Governance & Accountability (4 indicators)	B-	C+	C-	D	D	D	D	D+
Information Sharing (10 indicators)	A-	B	C	C	C-	D-	C-	D+
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

1.1 Key Findings

- **The top 3 companies from last time, Anthropic, OpenAI and Google DeepMind, hold their position, with Anthropic receiving the best score in every domain.** Anthropic has sustained its leadership in safety practices through consistently high transparency in risk assessment, a comparatively well-developed safety framework, substantial investment in technical safety research, and governance commitments reflected in its Public Benefit Corporation structure and support for state-level legislation such as SB 53. However, it also shows areas of deterioration, including the absence of a human uplift trial in its latest risk-assessment cycle and a shift toward using user interactions for training by default.
- **There is a substantial gap between these top three companies and the next tier (xAI, Z.ai, Meta, DeepSeek, and Alibaba Cloud), but recent steps taken by some of these companies show promising signs of improvement that could help close this gap in the next iteration.** The next-tier companies still face major gaps in risk-assessment disclosure, safety-framework completeness, and governance structures such as whistleblowing policies. That said, several companies have taken meaningful steps forward: Meta's new safety framework may support more robust future disclosures, and Z.ai has indicated that it is developing an existential-risk plan.

- **Existential safety remains the sector's core structural failure, making the widening gap between accelerating AGI/superintelligence ambitions and the absence of credible control plans increasingly alarming.** While companies accelerate their AGI and superintelligence ambitions, none has demonstrated a credible plan for preventing catastrophic misuse or loss of control. No company scored above a D in this domain for the second consecutive edition. Moreover, although leaders at firms such as Anthropic, OpenAI, Google DeepMind, and Z.ai have spoken more explicitly about existential risks, this rhetoric has not yet translated into quantitative safety plans, concrete alignment-failure mitigation strategies, or credible internal monitoring and control interventions.
- **xAI and Meta have taken meaningful steps towards publishing structured safety frameworks, although limited in scope, measurability, and independent oversight**. Meta introduces a relatively comprehensive safety framework with the only outcome-based thresholds, although its trigger for mitigation is set too high and decision-making authority remains unclear. Meanwhile, xAI has formalized its safety framework with quantitative thresholds, but it remains narrow in risk coverage and does not specify how threshold breaches translate into mitigation mechanisms.
- **More companies have conducted internal and external evaluations of frontier AI risks, although the risk scope remains narrow, validity is weak, and external reviews are far from independent.** Compared to the last edition, xAI and Z.ai both shared more about their risk assessment processes, joining Anthropic, OpenAI and Google DeepMind. However, reviewers have pointed out that disclosures still fall short: key risk categories are under-addressed, external validity is not adequately tested, and external reviewers are not truly "independent."
- **Although there were no Chinese companies in the Top 3 group, reviewers noted and commended several of their safety practices mandated under domestic regulation.** Domestic regulations, including binding requirements for content labeling and incident reporting, and voluntary national technical standards outlining structured AI risk-management processes, give Chinese firms stronger baseline accountability for some indicators compared to their Western counterparts.
- **Companies' safety practices are below the bar set by emerging standards, including EU AI Code of Practice.** Reviewers underscored the persistent gap between published governance frameworks and actual safety practices of companies across industry, noting that companies still fail to meet basic requirements such as independent oversight, transparent threat modeling, measurable thresholds, and clearly defined mitigation triggers

Taken together, these findings point to a frontier-AI ecosystem where companies' safety commitment continues to lag far behind its capability ambition. Even the strongest performers lack the concrete safeguards, independent oversight, and credible long-term risk-management strategies that such powerful systems demand, while the rest of the industry remains far behind on basic transparency and governance obligations. This widening gap between capability and safety leaves the sector structurally unprepared for the risks it is actively creating.

Note: the evidence was collected up until November 8, 2025 and does not reflect recent events such as the releases of Google DeepMind's Gemini 3 Pro, xAI's Grok 4.1, OpenAI's GPT-5.1, or Anthropic's Claude Opus 4.5.

1.2 Company Progress Highlights and Improvement Recommendations

All companies must move beyond high-level existential-safety statements and produce concrete, evidence-based safeguards with clear triggers, realistic thresholds, and demonstrated monitoring and control mechanisms capable of reducing catastrophic-risk exposure—either by presenting a credible plan for controlling and aligning AGI/ASI or by clarifying that they do not intend to pursue such systems.

Company	*Progress Highlights*	*Improvement Recommendations*
Anthropic	▪ Anthropic has increased transparency by filling out the company survey for the AI Safety Index. ▪ Anthropic has improved governance and accountability mechanisms by sharing more details about its whistleblower policy and promising to release a public version soon. ▪ Compared to other US companies, Anthropic has been relatively supportive of both international and U.S. state-level governance and legislative initiatives related to AI safety.	▪ **Make thresholds and safeguards more concrete and measurable** by replacing qualitative, loosely defined criteria with quantitative risk-tied thresholds, and by providing clearer evidence and documentation that deployment and security safeguards can meaningfully mitigate the risks they target. ▪ **Strengthen evaluation methodology and independence**, including moving beyond fragmented, weak-validity, task-based assessments and incorporating latent-knowledge elicitation, involving uncensored and credibly independent external evaluators.
OpenAI	▪ OpenAI has documented a risk assessment process that spans a wider set of risks and provides more detailed evaluations than its peers. ▪ Although OpenAI's new governance structure has been criticized, reviewers considered a public benefit corporation to be better than a pure for-profit corporation.	▪ **Make safety-framework thresholds measurable and enforceable**, by clearly defining when safeguards trigger, linking thresholds to concrete risks, and demonstrating proposed mitigations can be implemented in practice. ▪ **Increase transparency and external oversight**, by aligning public positions with stated safety commitments, and creating more and stronger open channels for independent audit. ▪ **Increase efforts to prevent AI psychosis and suicide,** and act less adversarially toward alleged victims. ▪ **Reduce lobbying against state-level regulations focused on AI safety.**
Google DeepMind	▪ Google DeepMind has improved in transparency by completing the AI Safety Index survey. ▪ Google DeepMind has improved governance and accountability mechanisms by sharing details about its whistleblower policy.	▪ **Strengthen risk-assessment rigor and independence**, by moving beyond fragmented and evaluations of weak validity, testing in more realistic noisy or adversarial conditions, and ensuring that external evaluators are not selectively chosen and compensated for. ▪ **Make thresholds and governance structures more concrete and actionable**, by defining measurable criteria, adapting Cyber CCLs to reflect volume-based risk, and establishing clear relationships with external governance, among internal governance bodies, and mechanisms for acting on thresholds being passed. ▪ **Increase efforts to prevent AI psychological harm and consider distancing itself from CharacterAI.** ▪ **Reduce lobbying against state-level regulations focused on AI safety.**
xAI	▪ xAI has formalized and published its frontier AI safety framework.	▪ **Improve breadth, rigor and independence of risk assessments**, including sharing more detailed evaluation methods and incorporating meaningful external oversight. ▪ **Consolidate and clarify the risk-management framework** with broader coverage of risk categories, measurable thresholds, assigned responsibilities, and defined procedures for acting on risk signals. ▪ **Allow more pre-deployment testing for future models than what was done for Grok4.**

Company	Progress Highlights	Improvement Recommendations
Z.ai	▪ Z.ai took a meaningful step toward external oversight, including allowing third-party evaluators to publish safety evaluation results without censorship and expressing willingness to defer to external authorities for emergency response.	▪ **Publicize the full safety framework and governance structure** with clear risk areas, mitigations, and decision-making processes. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on system and operational risks benchmarks, content-risk benchmarks and safety benchmarks. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ Consider signing the EU AI Act Code of Practice.
Meta	▪ Meta has formalized and published its frontier AI safety framework with clear thresholds and risk modeling mechanisms.	▪ **Improve breadth, depth and rigor of risk assessments and safety evaluations**, including clarifying methodologies as well as sharing more robust internal and external evaluation processes. ▪ **Strengthen internal safety governance by** establishing empowered oversight bodies, transparent whistleblower protections, and clearer decision-making authority for development and deployment safeguards. ▪ **Foster a culture that takes frontier-level risks more seriously**, including a more cautious stance toward releasing model weights. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts, signing the EU AI Act Code of Practice, and providing more substantive disclosures in the model card.
DeepSeek	▪ DeepSeek's employees have become more outspoken about frontier AI risks and the company has contributed to standard-setting for these risks.	▪ **Establish and publish a foundational safety framework and risk-assessment process**, including system cards and basic model evaluations. ▪ Establish and publish a whistle-blower policy and bug bounty program. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on benchmarks that evaluate system & operational Risks, content safety risks, societal risks, legal & rights-related risks, fairness, and safety. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts. ▪ Consider signing the EU AI Act Code of Practice.
Alibaba Cloud	▪ Alibaba Cloud has contributed to the binding national standards on watermarking requirements.	▪ **Establish and publish a foundational safety framework and risk-assessment process**, including system cards and basic model evaluations. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on truthfulness, fairness, and safety benchmarks. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts. ▪ Consider signing the EU AI Act Code of Practice.

1.3 Methodology

Index Structure: The Winter 2025 Index evaluates eight leading AI companies on 35 indicators spanning six critical domains. The eight companies include Anthropic, OpenAI, Google DeepMind, xAI, Z.ai, Meta, DeepSeek, Alibaba Cloud. The indicators are listed below, and more detailed definitions can be found in Section 3.1.

Risk Assessment

Internal Testing
- Dangerous Capability Evaluations
- Elicitation for Dangerous Capability Evaluations
- Human Uplift Trials

External Testing
- Independent Review of Safety Evaluations
- Pre-deployment External Safety Testing
- Bug Bounties for System Vulnerabilities

Current Harms

Safety Performance
- Stanford's HELM Safety Benchmark
- Stanford's HELM AIR Benchmark
- TrustLLM Benchmark
- Center for AI Safety Benchmarks

Digital Responsibility
- Protecting Safeguards from Fine-tuning
- Watermarking
- User Privacy

Safety Frameworks

Risk Identification

Risk Analysis and Evaluation

Risk Treatment

Risk Governance

Existential Safety

Existential Safety Strategy

Internal Monitoring and Control Interventions

Technical AI Safety Research

Supporting External Safety Research

Governance & Accountability

Company Structure & Mandate

Whistleblowing Protection
- Whistleblowing Policy Transparency
- Whistleblowing Policy Quality Analysis
- Reporting Culture & Whistleblowing Track Record

Information Sharing

Technical Specifications
- System Prompt Transparency
- Behavior Specification Transparency

Voluntary Commitment
- G7 Hiroshima AI Process Reporting
- EU General-Purpose AI Code of Practice
- Frontier AI Safety Commitments (AI Seoul Summit, 2024)
- FLI AI Safety Index Survey Engagement
- Endorsement of the Oct. 2025 Superintelligence Statement

Risks & Incidents
- Serious Incident Reporting & Government Notifications
- Extreme-Risk Transparency & Engagement

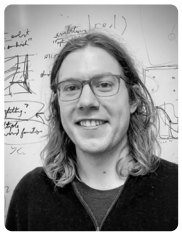
Public Policy
- Policy Engagement on AI Safety Regulations

Data Collection: The Index collected evidence up until November 8, 2025, combining publicly available materials—including model cards, research papers, and benchmark results—with responses from a targeted company survey designed to address specific transparency gaps in the industry, such as transparency on whistleblower protections and external model evaluations. Anthropic, OpenAI, Google DeepMind, xAI and Z.ai have submitted their survey responses. The complete evidence base is documented in Appendix A and Appendix B.

Expert Evaluation: An independent panel of eight leading AI researchers and governance experts reviewed company-specific evidence and assigned domain-level grades (A-F) based on absolute performance standards with discretionary weights. Reviewers provided written justifications and improvement recommendations. Final scores represent averaged expert assessments, with individual grades kept confidential.

1.4 Independent Review Panel

The scoring was conducted by a panel of distinguished AI experts:



David Krueger

David Krueger is an Assistant Professor in Robust, Reasoning and Responsible AI in the Department of Computer Science and Operations Research (DIRO) at University of Montreal, a Core Academic Member at Mila, and an affiliated researcher at UC Berkeley's Center for Human-Compatible AI, and the Center for the Study of Existential Risk. His work focuses on reducing the risk of human extinction from AI.

Dylan Hadfield-Menell

Dylan Hadfield-Menell is an Assistant Professor at MIT, where he leads the Algorithmic Alignment Group at the Computer Science and Artificial Intelligence Laboratory (CSAIL). A Schmidt Sciences AI2050 Early Career Fellow, his research focuses on safe and trustworthy AI deployment, with particular emphasis on multi-agent systems, human-AI teams, and societal oversight of machine learning.

Stuart Russell

Stuart Russell is a Professor of Computer Science at the University of California at Berkeley and Director of the Center for Human-Compatible AI and the Kavli Center for Ethics, Science, and the Public. He is a member of the National Academy of Engineering and a Fellow of the Royal Society. He is a recipient of the IJCAI Computers and Thought Award, the IJCAI Research Excellence Award, and the ACM Allen Newell Award. In 2021 he received the OBE from Her Majesty Queen Elizabeth and gave the BBC Reith Lectures. He coauthored the standard textbook for AI, which is used in over 1500 universities in 135 countries.

Sharon Li

Sharon Li is an Associate Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. Her research focuses on algorithmic and theoretical foundations of safe and reliable AI, addressing challenges in both model development and deployment in the open world. She serves as the Program Chair for ICML 2026. Her awards include a Sloan Fellowship (2025), NSF CAREER Award (2023), MIT Innovators Under 35 Award (2023), Forbes 30under30 in Science (2020), and "Innovator of the Year 2023" (MIT Technology Review). She won the Outstanding Paper Award at NeurIPS 2022 and ICLR 2022.

Jessica Newman

Jessica Newman is the Founding Director of the AI Security Initiative, housed at the Center for Long-Term Cybersecurity at the University of California, Berkeley. She serves as an expert in the OECD Expert Group on AI Risk and Accountability and contributes to working groups within the U.S. Center for AI Standards and Innovation, EU Code of Practice Plenaries, and other AI standards and governance bodies.

Sneha Revanur

Sneha Revanur is the founder and president of Encode, a global youthled organization advocating for the ethical regulation of AI. Under her leadership, Encode has mobilized thousands of young people to address challenges like algorithmic bias and AI accountability. She was featured on TIME's inaugural list of the 100 most influential people in AI.

Tegan Maharaj

Tegan Maharaj is an Assistant Professor in the Department of Decision Sciences at HEC Montréal, where she leads the ERRATA lab on Ecological Risk and Responsible AI. She is also a core academic member at Mila. Her research focuses on advancing the science and techniques of responsible AI development. Previously, she served as an Assistant Professor of Machine Learning at the University of Toronto.

Yi Zeng

Yi Zeng is an AI Professor at the Chinese Academy of Sciences, the Founding Dean of the Beijing Institute of AI Safety and Governance, and the Director of the Beijing Key Laboratory of Safe AI and Superalignment. He serves on the UN High-level Advisory Body on AI, the UNESCO Ad Hoc Expert Group on AI Ethics, the WHO Expert Group on the Ethics/Governance of AI for Health, and the National Governance Committee of Next Generation AI in China. He has been recognized by the TIME100 AI list.

2 Introduction

Frontier AI systems are now advancing with such speed and autonomy that make questions of near-term harms and long-term controllability increasingly salient. While today's AI systems already raise serious concerns around misuse and reliability, the development of more advanced, highly agentic, and self-improving models introduces risks at an entirely different scale and impact. As capabilities rise, both the opportunities offered by these systems and the risks they pose expand accordingly. Yet capability alone does not determine the overall risk landscape; it is also shaped by factors such as geopolitical competition, safety priorities, and public consensus. Because leading AI companies sit closest to these emerging thresholds, the safeguards they build—or fail to build—will heavily influence whether increasingly capable systems remain controllable or aligned with human intentions and values as they advance.

In response to this growing urgency, the AI Safety Index—developed by the Future of Life Institute together with an independent panel of experts in AI safety, governance, and technical evaluation—offers an independent assessment of how responsibly the world's leading AI companies are developing and deploying frontier systems. The Index evaluates companies safety practices on 35 indicators across six domains, from frontier risk management frameworks, to pre-deployment safety evaluations, from internal governance structure to external information sharing. By presenting results in a format accessible to both specialists and general audiences, the Index provides a transparent, evidence-based, and comparative picture of how companies manage risks as their systems become more capable, helping to identify where best practices are emerging and where critical gaps remain.

This iteration arrives at a moment when international expectations for corporate responsibility are becoming more concrete. New regulatory and governance initiatives, such as the G7 Hiroshima AI Process, the EU AI Code of Practice, California's SB53, and strengthened evaluation protocols from national AI Safety Institutes, are raising the baseline for what responsible behavior should look like. In this context, it is increasingly important to examine how companies are responding to these emerging obligations and voluntary commitments, and how these responses align with the scale of their stated ambitions for increasingly capable systems. The broader global consensus remains clear: rapidly advancing capabilities require urgent investment in alignment research and major improvements in risk-management practices.

Therefore, in this iteration, we evaluate eight frontier AI companies from across the world—including Anthropic, OpenAI, Google DeepMind, xAI, Z.ai, Meta, DeepSeek, and Alibaba Cloud—using a set of indicators that remain largely consistent with the previous edition. Keeping the indicators stable allows not only meaningful comparison across companies, but also comparison across iterations, making it possible to track how firms' safety practices evolve over time. This edition continues to serve as a practical and public-facing tool for tracking corporate behavior, identifying emerging best practices, and surfacing critical gaps in preparedness. By making companies' risk-management practices more visible and comparable, the Index aims to strengthen incentives for responsible development and narrow the gap between formal commitments and real-world actions, especially at a time when the stakes continue to rise.

3 Methodology

The AI Safety Index evaluates and grades the safety practices from AI companies in four steps: indicator selection, company selection, evidence collection, and grading.

3.1 Indicator Selection

To closely examine AI companies' safety practices throughout the lifecycle, we use 32 out of 34 indicators from the Summer 2025 edition, spanning six domains. The domains capture different aspects of responsible AI development and deployment, including risk assessment, current harms, safety framework, existential risk strategy, governance and accountability, as well as information sharing and public messaging, echoing principles embedded in regulatory obligations and voluntary commitments frameworks including the EU AI Code of Practice and the G7 Hiroshima Process. In particular, the Index highlights the existential risk strategy—a dimension not explicitly addressed in leading governance frameworks—because proactive planning for existential risk has become a pressing need, as emphasized by leading AI technical researchers and governance experts, including Bengio et al. (2024).

Two indicators from the original set, based on one-off robustness evaluations from UK's AI Safety Institute (AISI) and Cisco, were removed due to the lack of replicable evaluation protocols for the newly released frontier AI systems. Instead, we adopt the CAIS Safety Index, which aggregates performance across a range of open and ongoing evaluations, including deception, harmful behavior, overconfidence, jailbreak resistance, and bioweapon misuse. With support from CAIS, these benchmarks were run on the most recent models, ensuring consistency for comparison.

Additionally, three new indicators were added to the Information Sharing and Public Messaging domain to more comprehensively monitor company participation in key global voluntary commitments on safeguarding against frontier AI risks: the EU AI Code of Practice, the Frontier AI Safety Commitments at the AI Seoul Summit, and the October 2025 Superintelligence Statement issued by FLI.

Risk Assessment

This domain evaluates the rigor and comprehensiveness of companies' risk identification and assessment processes for their current flagship models. The focus is on implemented assessments, not stated commitments.

Group	Indicator Title	Summary
Internal testing	Dangerous Capability Evaluations	Tracks whether developers assess AI systems for harmful capabilities like cyber-offense, autonomous replication, or influence operations.
	Elicitation for Dangerous Capability Evaluations	Evaluates how transparently companies disclose and share their elicitation strategy used in dangerous capability evaluations.
	Human Uplift Trials	Evaluates whether companies conduct controlled experiments to measure how AI may increase users' ability to cause real-world harm.
External testing	Independent Review of Safety Evaluations	Assess whether third-party experts independently verify and critique the quality and accuracy of a developer's safety evaluations.
	Pre-deployment External Safety Testing	Measures whether independent, unaffiliated experts are given meaningful access to test a model's safety before public release.
	Bug Bounties for System Vulnerabilities	Assess whether developers offer structured incentives for discovering and disclosing safety issues specific to AI model behavior.

Current Harms

This domain covers demonstrated safety outcomes rather than commitments or processes. It focuses on the AI model's performance on safety benchmarks and the robustness of implemented safeguards against adversarial attacks.

Safety Performance	Stanford's HELM Safety Benchmark	Evaluates how language models perform on key safety metrics like robustness, fairness, and resistance to harmful behavior.
	Stanford's HELM AIR Benchmark	Measures AI model safety and security on benchmark aligned with emerging government regulations and company policies.
	TrustLLM Benchmark	Assesses a model's trustworthiness across dimensions such as safety, ethics, and alignment with human values and expectations.
	Center for AI Safety Benchmarks	Measures AI safety behaviors including resistance to misuse, appropriate refusals, calibration accuracy, honesty under pressure, and ethical restraint in scenarios.
Digital Responsibility	Protecting Safeguards from Fine-tuning	Evaluates whether AI providers implement protections that prevent fine-tuning from disabling important safety mechanisms or filters.
	Watermarking	Assess whether AI outputs are marked in a detectable way to help track origin and reduce misinformation or misuse.
	User Privacy	Measures the degree to which an AI company protects user data from extraction, exposure, or inappropriate use by models.

Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. This comprehensive analysis was conducted by the non-profit research organisation SaferAI.

Risk Identification	Evaluates whether companies systematically identify AI risks through comprehensive methods, including literature review, red teaming, and diverse threat modeling techniques.
Risk Analysis & Evaluation	Assesses whether companies translate abstract risk tolerances into concrete, measurable thresholds that trigger specific responses
Risk Treatment	Measures whether companies implement comprehensive mitigation strategies across containment, deployment safeguards, and affirmative safety assurance, with continuous monitoring throughout the AI lifecycle
Risk Governance	Examines whether companies establish clear risk ownership, independent oversight, safety-oriented culture, and transparent disclosure of their risk management approaches and incidents

Existential Safety

This domain examines companies' preparedness for managing extreme risks from future AI systems that could match or exceed human capabilities, including stated strategies and research for alignment and control.

Existential Safety Strategy	Assesses whether companies developing AGI publish credible, detailed strategies for mitigating catastrophic and existential AI risks, including alignment and control, governance, and planning.
Internal Monitoring and Control Interventions	Evaluates whether companies implement technical controls and protocols to detect and prevent model misalignment during internal use.
Technical AI Safety Research	Tracks whether companies publish research relevant to extreme-risk mitigation, including areas like interpretability, scalable oversight, and dangerous capability evaluations.
Supporting External Safety Research	Assesses the extent to which companies support independent AI safety work through mentorships, funding, model access, and collaboration with external researchers.

Governance & Accountability

This domain evaluates how openly companies share technical, safety, and governance information, and how their public and legislative messaging align with responsible AI governance

Company Structure & Mandate		Evaluates whether a company's legal and governance setup includes enforceable commitments that prioritize safety over profit incentives.
Whistleblowing Protections	Whistleblowing Policy Transparency	Assesses how publicly accessible and complete a company's whistleblowing system is, including reporting channels, protections, and transparency of outcomes.
	Whistleblowing Policy Quality Analysis	Rates the comprehensiveness and alignment of a company's whistleblowing policy with international best practices and AI-specific safety needs.
	Reporting Culture & Whistleblowing Track Record	Examines whether the company climate makes employees feel they can safely report AI safety concerns, based on leadership behavior, third-party evidence, and past incidents.

Information Sharing

This section gauges how openly firms share information about products, risks, and risk management practices. Indicators cover voluntary cooperation, transparency on technical specifications, and risk/incident communication.

Technical Specifications	System Prompt Transparency	Assesses whether companies publicly disclose the actual system prompts used in their deployed AI models, including version histories and design rationales.
	Behavior Specification Transparency	Evaluates if developers publish detailed and up-to-date documentation explaining their models' intended behavior, values, and decision-making logic across diverse scenarios.
Voluntary Cooperation	G7 Hiroshima AI Process Reporting	Tracks whether companies submitted detailed safety and governance disclosures to the G7 Hiroshima AI Process, reflecting their commitment to transparency.
	EU General-Purpose AI Code of Practice	Demonstrates AI companies' voluntary compliance with EU AI Act General-Purpose AI (GPAI) obligations by signing the non-binding guidelines.
	Frontier AI Safety Commitments (AI Seoul Summit, 2024)	Measures adherence to voluntary pledges by leading AI companies to develop safety frameworks for evaluating and managing severe AI risks.
	FLI AI Safety Index Survey Engagement	Reports which companies voluntarily completed and submitted FLI's detailed safety survey to supplement publicly available information.
	Endorsement of the Oct. 2025 Superintelligence Statement	Indicates whether a company has endorsed calls to prohibit superintelligence development until broad scientific consensus confirms safety and controllability.
Risks & Incidents	Serious Incident Reporting & Government Notifications	Evaluates public commitments, frameworks, and track records around reporting serious AI-related incidents to governments and peers.
	Extreme-Risk Transparency & Engagement	Measures whether company leaders publicly acknowledge catastrophic AI risks and proactively communicate those concerns to external audiences.
Public Policy	Policy Engagement on AI Safety Regulations	Tracks company involvement in shaping AI safety laws through public statements, consultations, testimony, and participation in regulatory coalitions.

3.2 Company Selection

The Index is primarily focused on companies that have deployed the most highly capable models currently available, or those that have previously done so and continue to invest actively in the development and deployment of new frontier systems. Based on the selection of Top 10 performing LLMs from LMArena's leaderboard overview as of October 8, 2025, this edition includes Anthropic, Google DeepMind, OpenAI, xAI, DeepSeek, Alibaba Cloud, and Z.ai[1]. Although Meta does not currently offer a model at the highest capability frontier, we are keeping it in the Index for one additional iteration in recognition of its sustained investment toward superintelligence-level research.

The flagship models that we evaluate are: Claude-Sonnet-4.5 (Anthropic), Gemini-2.5-Pro (Google DeepMind), GPT-5 (OpenAI), Grok-4 (xAI), R1 (DeepSeek), Qwen3-Max (Alibaba Cloud), and GLM-4.6 (Z.ai).

3.3 Related Work

Related Work: Several notable related efforts that drive transparency and accountability within the industry continue to inspire and complement the AI Safety Index. The most comprehensive of these efforts include SaferAI's in-depth analysis and ranking of AI companies' public safety frameworks (most recently updated as of October 2025), and two projects by Zach Stein-Perlman—AILabWatch.org (most recently updated as of September 15, 2025) and AISafetyClaims.org (most recently updated as of September 1, 2025)—which regularly provide detailed and technical evaluations of how leading AI companies work to avert catastrophic risks from advanced AI. Complementing these, the OECD report published in September 2025 synthesizes disclosures submitted through the G7's voluntary reporting framework and offers one of the first comparative, policy-grounded views of companies' governance and risk-management practices (Perset and Fialho Esposito, 2025). Earlier efforts include the Foundation Model Transparency Index in October 2023 and May 2024 published by Stanford Center for Research and Foundational Models (CRFM), which provides an empirical baseline for model transparency across the ecosystem.

Incorporated Work: Where appropriate, the 2025 Index incorporates existing comparative analysis led by credible research institutions.

In the Safety Framework domain, the Index draws on the indicator set developed by SaferAI's in-depth assessment of companies' published safety frameworks, while leaving all scoring to the independent reviewers convened by FLI. SaferAI is a leading governance and research non-profit with significant expertise in AI risk management.

The Index further integrates AILabWatch.org's tracking of technical AI safety research within the Existential Safety domain and complements it in two ways: by adding research published after the tracker's most recent update, and by incorporating safety-relevant research from companies not included in AILabWatch's coverage.

Our research on the quality of companies' whistleblowing policies in the 'Governance & Accountability' domain was enabled through support from OAISIS, a non-profit supporting individuals working at the frontier of AI who want to flag risks.

The 'Current Harms' domain evaluates flagship model performance on leading safety benchmarks, including the TrustLLM benchmark, the HELM AIR-Bench and HELM Safety benchmarks by Stanford's CRFM, and the Safety Index benchmarks curated by the Center of AI Safety (CAIS) AI Dashboard.

[1] The archived leaderboard on October 8, 2025 can be retrieved at this link: https://archive.ph/qvLY3.

3.4 Evidence Collection

The evidence collected for this iteration of the Index covers information up until November 8, 2025, drawing from publicly available information and a dedicated company survey for additional voluntary disclosures. Throughout the data collection process, FLI aimed to minimize bias and ensure a fair evaluation by applying consistent search protocols and evidence standards across companies.

To ensure fair evaluation across companies in China and those in the US and UK, this iteration introduces a methodological improvement that directly addresses the limitations identified last year. The Index now includes a concise, structured section explaining how China's regulatory system—across binding national laws, local regulations, voluntary technical standards, draft instruments, and policy guidance—shapes company behavior and disclosure practices. This addition enables reviewers to interpret Chinese companies' evidence within the regulatory environment they operate in, rather than through assumptions derived from US and UK contexts that emphasize voluntary self-governance and public documentation. By integrating this regulatory mapping into each relevant domain, the Index aims to improve cross-jurisdictional comparability and reduce systematic bias in grading.

In addition, this iteration incorporates a structured mapping to the EU AI Code of Practice. For each domain, we identify which commitments in the Code are most relevant and present them as a baseline reference for what voluntary obligations for many of the companies included currently look like. This mapping is provided solely as contextual material to help reviewers situate the indicators within emerging governance expectations; it does not prescribe grading thresholds, or function as an official rubric. Instead, graders are encouraged to use their own expert judgment, drawing on the EU AI Code of Practice as one of several reference points when interpreting companies' safety practices, particularly as firms navigate both compliance expectations and their own frontier-model development ambitions.

Desk research: Our evidence base primarily consists of public documentation that companies have released about their AI systems and risk management practices. This includes technical model cards detailing capabilities and limitations, peer-reviewed research papers on safety methodologies, official policy documents, blog posts outlining safety commitments, and recordings or transcripts of leadership interviews or testimony before government bodies. We further incorporated metrics of flagship model performance on external safety benchmarks, news reports from credible media outlets, and reports of relevant assessments by independent research organizations.

Company survey: To supplement public information, FLI created a 34-question survey that addresses current gaps in voluntary disclosures. The survey was sent out via e-mail on October 13, 2025 and firms were given until October 31, 2025 to respond. The survey can be reviewed in full in Appendix B. The survey questions have been kept the same from the Summer 2025 iteration in order to be more consistent and show changes over time. They specifically focus on risk management-related domains where current transparency standards in the industry are lacking, such as whistleblowing policies, external third-party model evaluations, and internal AI deployment practices. We received survey responses from five companies (OpenAI, xAI, Z.ai, Google DeepMind, Anthropic), representing 62.5% of assessed firms. Meta, DeepSeek, and Alibaba Cloud have not submitted a response.

Grading Sheets: The evidence collected for this edition of the Index was organized into the grading sheets presented in Appendix A. These sheets are divided across six domains and provide company-specific information for each of the 35 indicators included in the current edition. For every indicator, the grading sheets outline its scope, explain the rationale for its inclusion, and reference relevant literature with hyperlinks where appropriate. We prioritized primary sources directly from companies over secondary reporting wherever possible. Investigative

journalism played an important role by surfacing practices that companies have not publicly disclosed. Survey responses submitted by companies were incorporated and clearly highlighted within the relevant indicators. Each domain also includes a concise description of the corresponding Chinese regulatory environment. Where applicable, indicators are mapped to commitments in the EU AI Code of Practice to help situate them within emerging governance expectations.

3.5 Grading

The grading process was designed to ensure an impartial and qualified evaluation of the companies' performance across the selected indicators, based on expertise of individual reviewers in relevant fields. It features a review panel of distinguished independent experts who assess the company-specific evidence for their assigned indicators and assign domain-level grades that represent companies' performance within these domains.

Review Panel: To ensure that the Index scores rest upon authoritative judgements, FLI selected a group of eight leading independent experts to grade company performance on the set of indicators. Panel members were selected for their domain expertise and absence of conflicts of interest. Because the Index spans technical AI safety, governance, and policy, the panel brings together specialists across these areas and reflects broader geographic diversity from the previous iterations. The panel thus features both renowned machine learning professors who specialize in alignment and control, and governance experts from the academic and non-profit sectors. The composition of the panel remained largely consistent with the previous edition. We are grateful to Sharon Li and Yi Zeng for joining the panel as new members. The review panel is introduced at the beginning of this document.

Grading Phase: Grading sheets and survey results were shared with the review panel for evaluation on November 10, 2025, and the grading period ended on November 20, 2025. After reviewing the evidence, reviewers assigned letter grades (A+ to F) to each company per domain. For each grade assigned to individual companies, reviewers could provide brief justifications and recommendations. They were also able to provide domain-level comments when feedback applied to multiple firms or to explain their judgments. Not every reviewer graded every domain, but experts were assigned domains relevant to their area of expertise. Importantly, no fixed weighting was imposed across indicators within a domain. This approach allowed expert reviewers to apply their judgment in emphasizing aspects they deemed most critical. The grading sheets provided to reviewers further contained grading scales based on absolute performance standards rather than relative rankings, ensuring consistent expectations regardless of company size or geography. Final domain scores were calculated by averaging all reviewer grades for that domain, provided at least three panelists submitted an assessment. Overall grades were then derived by averaging the domain-level scores.

3.6 Limitations

Information Availability and Verification

Our evaluation relies primarily on public information, which creates fundamental constraints. Companies control what they disclose, despite occasional cases of whistleblowing, making it difficult to distinguish between poor transparency and poor strategy and implementation. We designed indicators around these transparency constraints, focusing where meaningful differences between companies were identifiable. For example, we cannot assess critical practices such as cybersecurity investments to protect model weights, as this information is rarely disclosed publicly but we instead look at how companies assess cybersecurity-related risks with their frontier AI systems.

The 35 indicators represent a subset of important practices for which meaningful evidence exists, but it does not comprehensively cover all safety dimensions. Furthermore, we cannot independently verify individual company claims and must assume official reports are truthful, which constitutes a significant limitation given the high stakes involved.

Alignment with Transparency Standards and Reporting Requirements

The transparency and disclosure expectations embedded across emerging governance instruments—ranging from voluntary codes such as the EU AI Code of Practice, to multilateral reporting frameworks like the G7 Hiroshima Process, to regulatory requirements such as California's SB 53—contain many overlapping elements but also differ substantially in scope, emphasis, and legal force. Incorporating every requirement would introduce unnecessary complexity, dilute the evaluative signal, and risk information fatigue among both expert reviewers and public audiences.

In this edition, we therefore focus on a limited and targeted mapping to the EU AI Code of Practice, using it only to provide contextual reference points for relevant indicators rather than as a comprehensive benchmarking standard. For future iterations, these governance instruments can help clarify which expectations should inform indicator design, highlight where existing rules set high or low bars, and expose gaps where critical safety practices remain unaddressed. At the same time, indicators covering high-stakes areas not yet reflected in current frameworks should continue to be emphasized through the AI Safety Index to ensure that it reflects where governance expectations fall short.

For policymakers, this alignment ultimately serves two purposes: showing how effectively existing rules shape company behavior, and identifying where further regulatory action or mandatory reporting may be most needed.

Methodological Constraints

Our focus on observable, documentable practices may undervalue crucial but hard-to-measure factors such as safety culture. Additionally, while we seek to diversify the grading panel with specialized expertise and geolocation focus, it cannot encompass all relevant domains across the companies that we review. Panelists' backgrounds inevitably shape their judgments, and there is an inherent tension between allowing experts to exercise domain-specific discretion in weighting indicators and maintaining full consistency across panelists and domains.

Moving Forward

We seek to address these limitations through continued refinement of our methods and closer engagement with policymakers, researchers, and practitioners who rely on the Index. Feedback from regulators and policy professionals is particularly valuable in helping us identify where clearer disclosure expectations, stronger reporting norms, or more precise indicator design would make the Index more actionable for real-world governance needs.

We will continue to document our sources, assumptions, and reviewer materials transparently, and we welcome constructive guidance on how to better incorporate hard-to-evaluate practices, reduce ambiguity in evidence interpretation, and strengthen cross-jurisdictional comparability. We encourage readers to share suggestions at policy@futureoflife.org and remain committed to advancing the Index with each iteration.

4 Results

Overall Rankings: Anthropic leads with a C+ (2.67), followed by OpenAI (C+, 2.31) and Google DeepMind (C, 2.08). The next group of companies cluster closely together, with xAI (D, 1.17), Z.ai (D, 1.12), Meta (D, 1.10), DeepSeek (D, 1.02), and Alibaba Cloud (D-, 0.98).Notably, no company scored above a C+, underscoring that even the strongest performers remain far from meeting adequate safety expectations.

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Overall Grade	C+	C+	C	D	D	D	D	D-
Score	2.67	2.31	2.08	1.17	1.12	1.10	1.02	0.98
Risk Assessment (6 indicators)	B	B	C+	D	D+	D	D	D
Current Harms (7 indicators)	C+	C-	C	F	D	D+	D+	D+
Safety Frameworks (4 indicators)	C+	C+	C+	D+	D-	D+	F	F
Existential Safety (4 indicators)	D	D	D	F	F	F	F	F
Governance & Accountability (4 indicators)	B-	C+	C-	D	D	D	D	D+
Information Sharing (10 indicators)	A-	B	C	C	C-	D-	C-	D+
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

4.1 Key Findings

- **The top 3 companies from last time, Anthropic, OpenAI and Google DeepMind, hold their position, with Anthropic receiving the best score in every domain.** Anthropic has sustained its leadership in safety practices through consistently high transparency in risk assessment, a comparatively well-developed safety framework, substantial investment in technical safety research, and governance commitments reflected in its Public Benefit Corporation structure and support for state-level legislation such as SB 53. However, it also shows areas of deterioration, including the absence of a human uplift trial in its latest risk-assessment cycle and a shift toward using user interactions for training by default.
- **There is a substantial gap between these top three companies and the next tier (xAI, Z.ai, Meta, DeepSeek, and Alibaba Cloud), but recent steps taken by some of these companies show promising signs of improvement that could help close this gap in the next iteration.** The next-tier companies still face major gaps in risk-assessment disclosure, safety-framework completeness, and governance structures such as whistleblowing policies. That said, several companies have taken meaningful steps forward: Meta's new safety framework may support more robust future disclosures, and Z.ai has indicated that it is developing an existential-risk plan.

- **Existential safety remains the sector's core structural failure, making the widening gap between accelerating AGI/superintelligence ambitions and the absence of credible control plans increasingly alarming.** While companies accelerate their AGI and superintelligence ambitions, none has demonstrated a credible plan for preventing catastrophic misuse or loss of control. No company scored above a D in this domain for the second consecutive issue. Moreover, although leaders at firms such as Anthropic, OpenAI, Google DeepMind, and Z.ai have spoken more explicitly about existential risks, this rhetoric has not yet translated into quantitative safety plans, concrete alignment-failure mitigation strategies, or credible internal monitoring and control interventions.
- **xAI and Meta have taken meaningful steps towards publishing structured safety frameworks, although limited in scope, measurability, and independent oversight**. Meta introduces a relatively comprehensive safety framework with the only outcome-based thresholds, although its trigger for mitigation is set too high and decision-making authority remains unclear. Meanwhile, xAI has formalized its safety framework with quantitative thresholds, but it remains narrow in risk coverage and does not specify how threshold breaches translate into mitigation mechanisms.
- **More companies have conducted internal and external evaluations of frontier AI risks, although the risk scope remains narrow, validity is weak, and external reviews are far from independent**. Compared to the last issue, xAI and Z.ai both shared more about their risk assessment processes, joining Anthropic, OpenAI and Google DeepMind. However, reviewers have pointed out that disclosures still fall short: key risk categories are under-addressed, external validity is not adequately tested, and external reviewers are not truly "independent."
- **Although there were no Chinese companies in the Top 3 group, reviewers noted and commended several of their safety practices mandated under domestic regulation. Domestic regulations**, including binding requirements for content labeling and incident reporting, and voluntary standards on model governance, give Chinese firms stronger baseline accountability for some indicators compared to their Western counterparts.
- **Companies' safety practices are below the bar set by emerging standards, including EU AI Code of Practice**. Reviewers underscored the persistent gap between published governance frameworks and actual safety practices of companies across industry, noting that companies still fail to meet basic requirements such as independent oversight, transparent threat modeling, measurable thresholds, and clearly defined mitigation triggers.

Taken together, these findings point to a frontier-AI ecosystem where companies' safety commitment continues to lag far behind its capability ambition. Even the strongest performers lack the concrete safeguards, independent oversight, and credible long-term risk-management strategies that such powerful systems demand, while the rest of the industry remains far behind on basic transparency and governance obligations. This widening gap between capability and safety leaves the sector structurally unprepared for the risks it is actively creating.

Note: the evidence was collected up until November 8, 2025 and does not reflect recent events such as the releases of Google DeepMind's Gemini 3 Pro, xAI's Grok 4.1, OpenAI's GPT-5.1, or Anthropic's Claude Opus 4.5.

4.2 Company Progress Highlights and Improvement Recommendations

All companies must move beyond high-level existential-safety statements and produce concrete, evidence-based safeguards with clear triggers, realistic thresholds, and demonstrated monitoring and control mechanisms capable of reducing catastrophic-risk exposure—either by presenting a credible plan for controlling and aligning AGI/ASI or by clarifying that they do not intend to pursue such systems.

Company	*Progress Highlights*	*Improvement Recommendations*
Anthropic	▪ Anthropic has increased transparency by filling out the company survey for the AI Safety Index. ▪ Anthropic has improved governance and accountability mechanisms by sharing more details about its whistleblower policy and promising to release a public version soon. ▪ Compared to other US companies, Anthropic has been relatively supportive of both international and U.S. state-level governance and legislative initiatives related to AI safety.	▪ **Make thresholds and safeguards more concrete and measurable** by replacing qualitative, loosely defined criteria with quantitative risk-tied thresholds, and by providing clearer evidence and documentation that deployment and security safeguards can meaningfully mitigate the risks they target. ▪ **Strengthen evaluation methodology and independence**, including moving beyond fragmented, weak-validity, task-based assessments and incorporating latent-knowledge elicitation, involving uncensored and credibly independent external evaluators.
OpenAI	▪ OpenAI has documented a risk assessment process that spans a wider set of risks and provides more detailed evaluations than its peers. ▪ Although OpenAI's new governance structure has been criticized, reviewers considered a public benefit corporation to be better than a pure for-profit corporation.	▪ **Make safety-framework thresholds measurable and enforceable**, by clearly defining when safeguards trigger, linking thresholds to concrete risks, and demonstrating proposed mitigations can be implemented in practice. ▪ **Increase transparency and external oversight**, by aligning public positions with stated safety commitments, and creating more and stronger open channels for independent audit. ▪ **Increase efforts to prevent AI psychosis and suicide,** and act less adversarially toward alleged victims. ▪ **Reduce lobbying against state-level regulations focused on AI safety.**
Google DeepMind	▪ Google DeepMind has improved in transparency by completing the AI Safety Index survey. ▪ Google DeepMind has improved governance and accountability mechanisms by sharing details about its whistleblower policy.	▪ **Strengthen risk-assessment rigor and independence**, by moving beyond fragmented and evaluations of weak validity, testing in more realistic noisy or adversarial conditions, and ensuring that external evaluators are not selectively chosen and compensated for. ▪ **Make thresholds and governance structures more concrete and actionable**, by defining measurable criteria, adapting Cyber CCLs to reflect volume-based risk, and establishing clear relationships with external governance, among internal governance bodies, and mechanisms for acting on thresholds being passed. ▪ **Increase efforts to prevent AI psychological harm and consider distancing itself from CharacterAI.** ▪ **Reduce lobbying against state-level regulations focused on AI safety.**
xAI	▪ xAI has formalized and published its frontier AI safety framework.	▪ **Improve breadth, rigor and independence of risk assessments**, including sharing more detailed evaluation methods and incorporating meaningful external oversight. ▪ **Consolidate and clarify the risk-management framework** with broader coverage of risk categories, measurable thresholds, assigned responsibilities, and defined procedures for acting on risk signals. ▪ **Allow more pre-deployment testing for future models than what was done for Grok4.**

Company	Progress Highlights	Improvement Recommendations
Z.ai	▪ Z.ai took a meaningful step toward external oversight, including allowing third-party evaluators to publish safety evaluation results without censorship and expressing willingness to defer to external authorities for emergency response.	▪ **Publicize the full safety framework and governance structure** with clear risk areas, mitigations, and decision-making processes. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on system and operational risks benchmarks, content-risk benchmarks and safety benchmarks. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ Consider signing the EU AI Act Code of Practice.
Meta	▪ Meta has formalized and published its frontier AI safety framework with clear thresholds and risk modeling mechanisms.	▪ **Improve breadth, depth and rigor of risk assessments and safety evaluations**, including clarifying methodologies as well as sharing more robust internal and external evaluation processes. ▪ **Strengthen internal safety governance by** establishing empowered oversight bodies, transparent whistleblower protections, and clearer decision-making authority for development and deployment safeguards. ▪ **Foster a culture that takes frontier-level risks more seriously**, including a more cautious stance toward releasing model weights. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts, signing the EU AI Act Code of Practice, and providing more substantive disclosures in the model card.
DeepSeek	▪ DeepSeek's employees have become more outspoken about frontier AI risks and the company has contributed to standard-setting for these risks.	▪ **Establish and publish a foundational safety framework and risk-assessment process**, including system cards and basic model evaluations. ▪ Establish and publish a whistle-blower policy and bug bounty program. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on benchmarks that evaluate system & operational Risks, content safety risks, societal risks, legal & rights-related risks, fairness, and safety. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts. ▪ Consider signing the EU AI Act Code of Practice.
Alibaba Cloud	▪ Alibaba Cloud has contributed to the binding national standards on watermarking requirements.	▪ **Establish and publish a foundational safety framework and risk-assessment process**, including system cards and basic model evaluations. ▪ **Substantially improve model robustness and trustworthiness** by improving performance on truthfulness, fairness, and safety benchmarks. ▪ **Establish and publicize a whistleblower policy** to enable employees to raise safety concerns without fear of retaliation. ▪ **Improve overall information sharing**, including by completing the AI Safety Index survey, participating in international voluntary standards efforts. ▪ Consider signing the EU AI Act Code of Practice.

4.3 Domain-level findings

Risk Assessment

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	B	B	C+	D	D+	D	D	D
Score	3.18	3.00	2.68	1.18	1.50	1.18	1.00	1.00

Anthropic, OpenAI, and Google DeepMind continue to lead on internal and external evaluations, with documented assessment processes reflected in their model cards and active bug bounty programs. Reviewers commended all three for including some well-designed internal experiments and strong capability-elicitation work, with OpenAI covering a broader set of risks and Anthropic providing relatively extensive bug bounty coverage. In addition, Z.ai was also recognized for its external evaluation practices, standing out as the only company that permits evaluators to publish results without censorship and for conducting external assessments before widespread internal deployment. By contrast, xAI and Meta provide much less detail in their model cards, although xAI offers a little more information on environment setup and quantitative benchmarks.

Despite these efforts, major gaps persist across the industry. No company has conducted Human Uplift Trials or secured truly independent reviews of safety evaluations. In addition, reviewers emphasized that companies do not meet the standards for independent oversight outlined in frameworks such as the EU AI Code of Practice. For example, companies including Anthropic, OpenAI, Google DeepMind and xAI acknowledge that external evaluators face restrictions on what can be published before deployment, while Google DeepMind further noting that its external evaluators are financially compensated by the company.

Moreover, the scope of risk assessments also remains narrow: while some serious harms appear moderately controlled, many significant risk categories remain unexamined. One reviewer pointed out no company assesses climate-related or environment-related risks, despite widespread controversy about data centers polluting water and harming local ecosystems. No company has published quantitative estimates of the probability that the AGI/superintelligence they aspire to build will be critically misaligned or escape control. Even heavily studied areas across the industry such as biorisk rely on task-specific testing with little work on latent-knowledge elicitation, adversarial conditions, or real-world deployment contexts.

Current Harms

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	C+	C-	C	F	D	D+	D+	D+
Score	2.43	1.77	2.10	0.57	1.00	1.33	1.67	1.43

Companies consistently scored poorly on current harm evaluations, although none of the tested models failed outright. Reviewers emphasized that "frequent safety failures, weak robustness, and inadequate control of serious harms are universal patterns," with uniformly low performance on trustworthiness benchmarks such

as truthfulness, fairness, and harmful-content generation. One reviewer argued that passing these tests should be considered unit tests of basic functionality,"yet model behavior frequently fell short. Moreover, one reviewer cautioned that benchmark results should be interpreted "with a grain of salt," as they are narrow, sometimes gamed, and may not reflect real-world risk, implying that true safety levels are likely lower than what the measurements reflect.

Anthropic scored highest in this domain while xAI performed the worst. Anthropic has consistently scored the highest across the benchmarks selected for this safety index while xAI has performed exceptionally poorly for the HELM AIR Benchmark. Meanwhile, companies also failed to uphold privacy principles seriously, as reviewers highlighted that "all models train on data from [user] interactions [by default]," a practice that exposes users to significant risks because sensitive information shared during interactions can later be retrieved. Unfortunately, Anthropic—previously the only company that did not use interaction data for training by default—shifted its policy in August 2025, contributing to its lower score.

A few companies received positive recognition for watermarking. Chinese companies comply with the binding national standards that require both explicit and implicit watermarking, and reviewers commended this baseline safeguard. Google was also praised for its watermarking practices, though one reviewer expressed concerns about its "decision to not make them user-accessible."

Safety Frameworks

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	C+	C+	C+	D+	D-	D+	F	F
Score	2.65	2.55	2.45	1.50	0.88	1.62	0.55	0.55

Anthropic, Google DeepMind, Meta, OpenAI, and xAI have all published safety frameworks, with Anthropic, Google DeepMind, and OpenAI offering the most structured approaches. These three outline risk areas, qualitative thresholds, and mitigation measures. In particular, Google DeepMind's framework is commended by the reviewer for expanding its framework to include harmful manipulation and misalignment risks and introducing early-warning evaluations to maintain a safety buffer. On the other hand, Meta's framework is notable for its use of outcome-based thresholds and clearer risk-modeling detail, though its risk coverage remains narrow and its governance pathways for halting development are undefined. xAI's framework, while containing certain quantitative thresholds, is criticized for having narrow risk coverage and thresholds that do not clearly influence deployment decisions.

DeepSeek, Z.ai and Alibaba do not have any published safety framework, and therefore received failing marks. However, reviewers acknowledge Zai's investment in its safety team and commitment to disclosing system prompts to regulators when "safety testing determines a model exceeds its "unacceptable-risk" threshold."

Even among companies with more structured safety frameworks, significant gaps remain. Thresholds are typically qualitative, vague, or not tied to measurable risk, and most frameworks focus narrowly on a limited set of categories such as CBRN risks while leaving major areas of systemic, societal, and alignment risk unexamined. Safeguards for deployment and security are often described only at a high level or as illustrative examples, with little evidence of concrete procedures or implementation. Engagement with external governance is also

limited, with few mechanisms for independent oversight. Among U.S.- and U.K.-based firms, only Anthropic includes explicit transparency commitments—such as inviting expert input and notifying U.S. authorities above ASL-2—and incorporates independent audit provisions.

In addition, reviewers raised concerns about the concentration of decision-making authority in senior leadership. At OpenAI, development and deployment decisions ultimately rest with company leadership, even though the Safety Advisory Group provides recommendations and the Board's Safety and Security Committee offers oversight. For Google DeepMind, the framework references several internal bodies that review and approve actions when alert thresholds are reached, but it remains unclear what respective expertise these groups have, how decisions are made, and whether any of them have the authority to halt deployment independent of executive leadership.

Existential Safety

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	D	D	D	F	F	F	F	F
Score	1.22	1.00	1.15	0.40	0.33	0.33	0.00	0.00

Companies performed most poorly across this domain. Reviewers emphasized that even the strongest performers make questionable assumptions in their existential-safety strategies, noting that firms are actively engaged in an explicit AGI race while lacking credible plans for preventing catastrophic misuse or loss of control—an inconsistency characterized as a "foundational hypocrisy." Although leadership at companies including Anthropic, Google DeepMind, OpenAI, xAI, and Z.ai have publicly expressed concerns about existential risk, the absence of concrete and actionable strategies was argued by a reviewer to render their self-assessments for the preparedness for existential risk "suspect at best."

Three companies—Anthropic, Google DeepMind, and OpenAI—score notably better than others, largely because they publish more safety research and outline higher-level risk-management frameworks for frontier-AI risks. For example, Google DeepMind's updated framework now takes into consideration scheming risks, and all three companies signal commitments to monitoring and control. However, these commitments lack binding safeguards. And reviewers argue that in some cases, the thresholds for intervention are set too high—for instance, Anthropic requires models to "automate the work of junior researchers" before certain mitigations would be triggered, and one reviewer noted that it is not realistic if "we are not good enough at eliciting and making effective use of AI capabilities."

Most other companies have shown little progress since the last iteration. xAI and Meta lack any commitments on monitoring and control despite having risk-management frameworks, and have not presented evidence that they invest more than minimally in safety research. DeepSeek, Alibaba Cloud, and Z.ai lack publicly available documents about existential safety strategy, although though reviewers positively noted that Z.ai is developing a plan for managing existential risk, and that it has disclosed a relatively concrete monitoring and control mechanism, even if it remains inadequate for "powerful scheming AI."

Finally, two reviewers raised broader concerns about whether releasing model weights— justified currently in the domain as supporting external safety research—might generate more potential harm than benefit in the context of frontier systems.

Governance & Accountability

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	B-	C+	C-	D	D	D	D	D+
Score	2.82	2.54	1.80	1.20	1.08	1.28	1.14	1.34

Anthropic, OpenAI, and Google DeepMind lead the field in governance and accountability. Anthropic and OpenAI operate as public benefit corporations that balance their mission and commercial success. One reviewer explicitly pointed out that although OpenAI's controversial restructuring is "not optimal," the final arrangement is still "far better than expected" as it retains the non-profit part as OpenAI Foundation and the for-profit arm has become a public benefit corporation (PBC) called "OpenAI Group PBC," with the non-profit still having significant control over the for-profit. OpenAI is also the only company with a public-facing whistleblowing policy, while Anthropic has indicated it plans to publish one soon. Google DeepMind, despite not having a public whistleblowing policy, received the highest overall governance-quality assessment based on its detailed disclosures in the survey response.

Other for-profit companies fall notably behind. Meta has no public whistleblowing policy, though its Code of Conduct offers limited transparency about internal practices and references a third-party–run Integrity Line. Similarly, Alibaba Cloud does not have a public whistleblowing policy, but its Code of Ethics indicates that employees have established channels for reporting concerns. While xAI also lacks a public whistleblowing policy, it provided more detail in its survey response to FLI, describing the internal role responsible for overseeing whistleblowing, the independence of investigations, the policy's scope, and protections related to confidentiality, non-retaliation, and reporting mechanisms. By contrast, Z.ai and DeepSeek do not publicly disclose any whistleblowing policy or its contents.

Information Sharing

	Anthropic	OpenAI	Google DeepMind	xAI	Z.ai	Meta	DeepSeek	Alibaba Cloud
Domain Grade	A-	B	C	C	C-	D-	C-	D+
Score	3.74	3.00	2.28	2.20	1.92	0.85	1.74	1.54

Company performance in information sharing and public messaging varies widely, with Anthropic and OpenAI leading while Meta and Alibaba Cloud lagging behind. Anthropic has been notably more transparent and more supportive of state-level AI safety regulations than its counterparts—publishing its system prompts, submitting HAIP compliance reports, signing the CoP, and publicly endorsing SB 53. OpenAI similarly releases and regularly updates detailed model specifications and engages actively in international voluntary transparency and safety efforts, though its score is reduced due to direct and indirect opposition to legislative proposals such as SB 1047 and support of federal preemption of state AI acts.

Meta received the lowest score in this domain, driven by leadership's track record of public dismissing AI existential risks and its aggressive lobbying against key regulatory initiatives, including SB 1047, the EU

AI Act, and the New York RAISE Act. Chinese companies also scored comparatively low, largely due to limited public communication by both company leadership and corporate channels. Nonetheless, reviewers acknowledged Z.ai's leadership's public endorsement of FLI's superintelligence statement, commended DeepSeek and Alibaba Cloud for their contributions to national AI-safety standards, and acknowledged that Chinese regulations impose mandatory incident-reporting obligations that strengthen baseline transparency.

Google DeepMind and xAI show a mix of strong and weak practices. Reviewers noted that Google's public messaging on AI safety is inconsistent: its leadership frequently speaks about extreme risks, and the company participates in international governance efforts such as HAIP and the CoP, yet in the U.S. it has opposed or lobbied against frontier AI legislation, including SB 53 and SB 1047. On the other hand, xAI, whose CEO Elon Musk speaks openly and regularly about extreme risks, released its system prompt after two controversial incidents involving alleged unauthorized changes. However, it engages less with international transparency commitments, didn't submit a HAIP report, and signed only the safety and security chapter of the CoP while opting out of sections on transparency and copyright.

5 Conclusions

The landscape that emerges from the Winter 2025 iteration of the Index is one marked by both expanding commitments and widening disparities between top tiers and lower tiers companies. New voluntary and binding frameworks—ranging from the EU AI Code of Practice to California's SB 53—have begun to set clearer expectations for safety practice disclosure, oversight, and risk governance. At the same time, more companies are pushing into the frontier tier of capabilities, accelerating a competitive dynamic that raises the stakes for robust and credible safeguards across the world. Yet even as companies advance their safety practices, the strongest performers still fall short of several important practices expected under these emerging frameworks, from genuinely independent evaluation to measurable risk thresholds, while the lower tier companies continue to fall short on basic elements such as safety frameworks, governance structures, and comprehensive risk assessment.

Still, this year's Index offers limited grounds for optimism. Several companies have made visible progress since the last iteration, particularly in increasing transparency and formalizing and publishing their safety frameworks. These steps signal an emerging alignment around baseline norms. But the improvements remain incremental, and they do not close the widening structural gap between capability development and safety preparedness. As international standards and regulatory reporting obligations continue to mature, companies should treat these frameworks not as aspirational guides but as anchor points for a decisive shift toward rigorous, enforceable, and independently validated safety measures. As one reviewer cautioned:

> *"Overall, companies generally are doing poorly, and even the best are making questionable assumptions in their safety strategies.*

The Future of Life Institute remains committed to tracking these critical developments through regular Index updates. We will continue working with our expert review panel and partner organizations to refine our assessments and highlight both concerning gaps and emerging best practices.

Bibliography

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari et al. "Managing extreme AI risks amid rapid progress." *Science* 384, no. 6698 (2024): 842-845.

Bommasani, Rishi, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. *The Foundation Model Transparency Index v1.1: May 2024*. Stanford Center for Research on Foundation Models, 2024. https://crfm.stanford.edu/fmti/paper.pdf.

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. "The Foundation Model Transparency Index." arXiv preprint arXiv:2310.12941, 2023. https://arxiv.org/abs/2310.12941.

Huang, Yue, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, and Yixuan Zhang, et al. "TrustLLM: Trustworthiness in Large Language Models." arXiv preprint arXiv:2401.05561, 2024. https://arxiv.org/abs/2401.05561.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, and Ananya Kumar, et al. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW.

Perset, Michel, and Sara Fialho Esposito. How Are AI Developers Managing Risks? *Insights from Responses to the Reporting Framework of the Hiroshima AI Process Code of Conduct*. Paris: OECD, 2025. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/09/how-are-ai-developers-managing-risks_fbaeb3ad/658c2ad6-en.pdf.

Phan, Long, and Dan Hendrycks. *CAIS AI Dashboard*. Center for AI Safety, 2025. https://dashboard.safe.ai.

SaferAI. "Risk Management Ratings — Frontier AI Companies." *SaferAI*. Accessed on October 3, 2025. https://ratings.safer-ai.org/

Stein-Perlman, Zach. *AILabWatch: Tracking AI Lab Research Outputs*. Accessed on October 3, 2025. https://ailabwatch.org/.

Stein-Perlman, Zach. *AISafetyClaims.org: Tracking Safety Claims Made by AI Companies*. Accessed on October 3, 2025. https://aisafetyclaims.org/.

Zeng, Yi, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. "AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies." Preprint, 2024. https://arxiv.org/abs/2407.17436v2.

Appendix A: Grading Sheets

Each of our panellists were presented with the full contents of this appendix to inform their grading decisions.

The grading sheets are broken down by domain, and panellists were asked to provide grades for each company per domain. Within each domain is a set of indicators: a collection of facts about the companies.

Grading Sheets

Risk Assessment

6 indicators

Current Harms

7 indicators

Safety Frameworks

4 indicators

Governance & Accountability

4 indicators

Existential Safety

4 indicators

Information Sharing and Public Messaging

10 indicators

Additional context on Chinese Regulatory System

How does it influence Chinese companies' behavior?

It is challenging to provide a fair comparison between frontier AI companies in China and those in the United States because of differing contexts. It is not obvious whether companies are more likely to abide by their own voluntary commitments (which are common in the U.S.) or draft laws and government standards that have not yet come into force (which are common in China). To enable our reviewers to draw their own conclusions, we will summarize the status of relevant Chinese laws and standards for each indicator.

In China, national and local regulations carry immediate force, as they carry legal and market-access consequences. Voluntary standards, while not legally binding, often serve as practical compliance references and are widely adopted in practice. Even draft regulations and policy guidance—at both national and local levels—may shape expectations and signal future directions, prompting companies to align early in order to sustain legitimacy and regulatory goodwill. In this context, the relative scarcity of voluntary safety commitments by Chinese companies may at least in part reflect differences in regulatory expectations and channels for policy engagement.

Below is a high-level summary of how each type of legislation or policy documents influence Chinese AI companies' behaviors.

National Binding Instruments

Binding national laws, regulations, and standards are legally enforceable instruments issued by the National People's Congress (NPC), the State Council, or ministries such as the Cyberspace Administration of China (CAC), the Ministry of Industry and Information Technology (MIIT), and the State Administration for Market Regulation (SAMR).

Current regulations include the Provisions on the Administration of Algorithmic Recommendation in Internet Information Services (2022), the Provisions on the Administration of Deep Synthesis of Internet Information Services (2022), the Interim Measures for the Management of Generative Artificial Intelligence Services (2023). The only binding standard is the National Standard on AI-generated content labeling and watermarking (2025).

These instruments carry direct legal force and set the non-negotiable baseline for companies if they want market access, and therefore companies comply immediately to avoid suspension, fines, or license revocation.

Local Binding Instruments

These are legally enforceable instruments enacted by provincial or municipal People's Congresses and implemented by local governments. Enforcement is carried out by local CAC, MIIT, or market-regulation branches.

Prominent examples include The Regulation of the Shanghai Municipality on Promoting the Development of the Artificial Intelligence Industry ("Shanghai Regulation", 2022), Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone ("Shenzhen Regulation", 2022). It is important to note that in Zhejiang—where Alibaba's Qwen models are registered with the provincial CAC office—and in Beijing—where Zhipu AI's GLM models are filed—there exist only local administrative regulations that govern how government agencies implement and enforce national AI directives, rather than local laws enacted by people's congresses that would impose binding obligations on enterprises. The binding rules shape behavior through localized incentives and compliance gatekeeping—firms align to secure compute resources, tax benefits, or pilot participation. Provincial or municipal measures could inform future actions on the national level.

Voluntary Technical Standard

These include GB/T (recommended national standards), industry standards, and local standards developed by technical committees such as TC260 (National Information Security Standardization Technical Committee) under the The Standardization Administration of China (SAC).

Prominent examples in the field of AI safety include GB/T 42888-2023 on Information Security Technology - Assessment Specification for Security of Machine Learning algorithms, GB/T 43191–2025 Basic Security Requirements for Generative AI Services, and GB/T 46347-2025 Artificial Intelligence - Risk Management Capability Management ("Risk Management Standard")

Companies adopt these standards voluntarily. However, in practice, these non-binding standards exert procedural and reputational pressure. Companies adopt them preemptively to pass CAC assessments and demonstrate compliance when filing with the government. Their influence is widespread but softer than law as they shape engineering, documentation, and testing practices without formal penalties.

Draft Regulations and Standards

They are typically issued by ministries such as the CAC, MIIT, TC 260 or municipal governments, as part of the government's legislative or standard-making agenda.

Prominent examples include Shanghai Draft Standard for Multimodal Model Safety Assessment ("Shanghai Draft", 2025).

Their influence is anticipatory and strategic: although not legally enforceable yet, they serve as early compliance signals, given that most are expected to be enacted with only limited modifications.

Strategic and Policy Guidance Documents

These include guidance documents issued by ministries or technical bodies (e.g., MOST, TC260), high-profile political speeches or party directives from senior leadership, and regulations or standards under drafting. While not enforceable, they shape the ideological framing for policymaking.

Prominent examples include Ethical Norms for New Generation Artificial Intelligence (MOST, 2021), Xi Jinping's 2024 speech emphasizing AI controllability, the AI Safety Governance Framework 1.0 (2024) and 2.0 (2025) by TC260 that introduces the plan to develop national AI safety standards and risk taxonomies, and Global AI Governance Action Plan (CAC, 2025).

These high-level policy guidance functions as a behavioral steering tool, compelling platform firms to anticipate regulatory trends, publicly align with state priorities, and adjust business practices long before formal laws are enacted.

Domain

Risk Assessment

This domain evaluates the rigor and comprehensiveness of companies' risk identification and assessment processes for their current flagship models. The focus is on implemented assessments, not stated commitments.

Table of Contents

Chinese Regulatory System Summary

At present, no binding national regulations or standards—whether mandatory or recommended—explicitly address frontier AI risks or define corresponding risk assessment processes. The Shanghai Draft offers early compliance guidance but its final scope, adoption timeline if adopted at the national level, and extrajurisdictional applicability remain uncertain. Nonetheless, the AI Safety Governance Framework 2.0 signals the government's intent to establish national standards to systematically address frontier risks in the near future.

Local Binding Instruments

Shenzhen Regulation (2022) requires high-risk AI applications to adopt a regulatory model of ex-ante assessment and risk warning (Article 66), although it doesn't specify which risks the service providers should assess. This does not apply to Z.ai's GLM models (Beijing) or Deepseek's R1 model (Zhejiang), and Alibaba's Qwen models (Zhejiang).

Draft Regulations and Standards

Article 5.8 of the Shanghai Draft enumerates potential high-risk capabilities of large models, including generation of malicious software, enabling the development of biological or chemical weapons, engaging in deceptive behavior, and exhibiting self-replication or self-improvement tendencies. However, Article 6.1.8 narrows the focus to cyber-related risks, requiring evaluation of the model's potential to uplift cyberattacks—specifically through the generation of malicious code, phishing emails, password cracking, vulnerability exploitation, and social-engineering attacks.

Article 7 of the Shanghai Draft covers three main aspects: evaluation methods, evaluation procedures, and reporting requirements. For methods, it outlines distinct evaluation approaches for text, image, voice, and video generation. For procedures, it specifies four key steps: establishing an evaluation committee, determining the scope and content of evaluation, conducting the evaluation work, and producing the final evaluation report. For reporting, it requires detailed documentation of methodologies (including automated testing, manual review, and user feedback mechanisms), analysis of false negatives and false positives, and concrete improvement suggestions. The final report must include both quantitative data and illustrative materials such as diagrams and case studies.

Strategic and Policy Guidance Documents Drafting in process: Artificial intelligence—Large language model alignment capability evaluation.

The AI Safety Governance Framework 2.0

Article 5.8 calls for the establishment of an AI safety evaluation system that integrates model and algorithm safety testing, general application safety testing, and scenario-specific safety testing.

Article 3.2.3 (c) explicitly calls for focusing on risk including loss of control over knowledge and capabilities of nuclear, biological, chemical, and missile weapons.

Specifically, Article 6.1.9 recommends regular safety evaluations and testing where a risk classification, grading, and optimization mechanism is established, clearly defining testing objectives, scope, and safety dimensions before each evaluation. It calls for the development of diverse testing datasets that cover a wide range of application scenarios, and the formulation of targeted model optimization strategies for different categories of risks.

Moreover, Article 5.11 calls for building global consensus and coordination mechanisms to address AI loss-of-control risks. It emphasizes strengthening end-use management of AI systems by setting specific safeguards for their application in nuclear, biological, chemical, and missile-related domains to prevent misuse. The clause promotes the adoption of trusted AI principles that integrate technical, ethical, and managerial dimensions, aiming to foster broad international alignment on responsible AI governance. It also requires developers to conduct regular testing to assess whether their models may pose potential technical loss-of-control risks.

Internal

Indicator

Dangerous Capability Evaluations

Definition

This indicator assesses whether organizations conduct systematic evaluations of dangerous capabilities before deploying frontier models. Priority domains include biological and chemical weapons, offensive cyber operations, recursive self-improvement risks, and behaviors associated with goal misalignment or deception. Evidence is drawn from model cards detailing testing methodologies and results. The focus is on external deployments, as there is insufficient transparency on internal deployments.

Why This Matters

Systematic evaluations for high-risk capabilities reflect institutional responsibility for managing low-probability, high-impact harms. In contrast to more routine risks—where market forces often suffice—frontier threats require deliberate foresight. Firms that fail to test for these dangers risk contributing to unmanaged systemic vulnerability.

EU AI Code of Practice (Safety and Security)

Measure 3.2 (Appendix 3.1)

Signatories will conduct at least state-of-the-art model evaluations in the modalities relevant to the systemic risk to assess the model's capabilities, propensities, affordances, and/or effects.

Model evaluations should be designed and conducted using methods that are appropriate for the model and the systemic risk and should include open-ended testing of the model.

Examples of model evaluation methods include: Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials.

The evaluation should ensure 1) internal validity, 2) external validity, 3) reproducibility. (Appendix 3.1)

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Biosecurity & Chemical Risk							
Final rounds of safety evaluations were conducted on the same model version that was released. Evaluations prioritize biological risks and do not conduct internal or external evaluations for chemical risk. **Safety Framework Classification** Evaluations test AI Safety Level 3 (ASL-3) and ASL-4 capability thresholds for related risks under Anthropic's Responsible Scaling Policy (RSP). **Evaluations scope** covers: 1) ASL-3: testing whether models can assist low-expertise actors in performing core biological threat workflows - Long-form virology tasks (task-based agentic evaluations co-developed with SecureBio, Deloitte, and Signature Science), - Multimodal virology (SecureBio VCT), - DNA Synthesis Screening Evasion (SecureBio) - LAB-Bench subset (expert-level biological skills assessment developed by FutureHouse) 2) ASL-4: testing whether models could substantially accelerate advanced or state-scale biological R&D - Creative biology (SecureBio) - Short-horizon computational biology tasks (Faculty.ai) **Methodological Details include:** 1) Environment and elicitation setup (e.g. containerization, tool integration, agent harness, "helpful-only" model variants, extended thinking mode etc.) 2) Human/AI baselines 3) Quantitative evaluation metrics (e.g. Rule-in/out thresholds, human & model baselines) System Card (pp. 125-136)	*Final rounds of safety evaluations were conducted on the same model version that was released.* Evaluations prioritize biological capability evaluations. **Safety Framework Classification** GPT-5 is treated as High capability in the Biological and Chemical domain under OpenAI's Preparedness Framework. **Evaluation Scope** covers: (1) Long-form biorisk questions (five stages of biothreat creation—ideation to release) (2) Multimodal virology troubleshooting (SecureBio/Center for AI Safety) (3) ProtocolQA open-ended troubleshooting (adapted from FutureHouse [Laurent et al., 2024]) (4) Tacit knowledge & troubleshooting (Gryphon Scientific, not published) (5) TroubleshootingBench focusing on real-world, experience-grounded wet-lab errors (6) Virology capabilities, human pathogen capabilities, molecular biology capabilities, world class biology (external evaluation by SecureBio) **Methodological Details** include: (1) Elicitation setup (e.g. maximum verbosity) (2) Human and expert baselines (3) Quantitative evaluation metrics System Card (pp. 23-27)	Evaluations have covered biological, chemical, nuclear, and radiological capabilities. **Safety Framework Classification** CBRN risks are tested for Uplift Level 1, with additional "alert-threshold" monitoring for early-warning signs of dangerous dual-use capabilities. It remains below the alert threshold. **Evaluation scope** includes: (1) Multiple choices quantitative questions: i) SecureBio VMQA4 single-choice; ii) FutureHouse LAB-Bench presented as three subsets (ProtocolQA, Cloning Scenarios, SeqQA) (Laurent et al., 2024); and iii) Weapons of Mass Destruction Proxy (WDMP) presented as the biology and chemistry data sets (Li et al., 2024). (2) Open-ended questions: qualitative assessment on knowledge-based, adversarial, and dual-use content in the biological, radiological and nuclear domains led by domain experts. **Methodological Details** include: (1) Quantitative and qualitative evaluation metrics (2) Human, expert, and model performance baselines System Card (pp. 12-14)	The system card mentions that Meta has conducted expert-designed and other targeted evaluations designed to assess whether the use of Llama 4 could meaningfully increase the capabilities of malicious actors to plan or carry out attacks using these types of weapons, however, no safety framework classification, methodological details and scope information are disclosed.	*Final rounds of safety evaluations were conducted on the same model version that was released.* Evaluations prioritize biological capability evaluations. **Safety Framework Classification** None **Evaluation Scope** covers: (1) Dual-use knowledge for bioweapons (2) Chemical knowledge **Methodological Details** include: (1) Benchmarks (WMDP Bio, WMDP Chem, BioLP-Bench, VCT [text-only]) (2) Quantitative metrics System Card (pp. 5)	**Not Mentioned**	*Final rounds of safety evaluations were conducted on the same model version that was released.* **Not Mentioned**	**Not Mentioned**

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Cybersecurity Risks							
Yes **Safety Framework Classification** **Ongoing assessment without formal threshold in RSP at any ASL.** **The Evaluation Scope** covers 1) General Cyber Evaluations - Quantitative results on CyberGym/Cybench - Anecdotal observations on triage and patching 2) Advanced Risk Evaluations - Irregular Challenges (23 private CTFs co-developed with Irregular to measure ability to discover and exploit complex vulnerabilities across categories including Web, Crypto, Pwn, Rev, Network) - Incalmo Cyber Ranges (25–50 hosts; co-developed with Carnegie Mellon University to test the model's capacity for long-horizon, multi-host cyber operation). **Methodological Details include** (1) Environment and elicitation (e.g. Kali-based sandbox, access to terminal, code editor, and standard penetration-testing tools) (2) Benchmarks and model performance baselines (3) Quantitative evaluation metrics System Card (pp. 32-45, 148)	**Yes** **Safety Framework Classification** Cyber capabilities are tracked as part of ongoing safety monitoring. **The Evaluation Scope** covers (1) Capture-the-Flag (CTF) Challenges across Web Application Exploitation, Reverse Engineering, Binary & Network Exploitation (pwn), Cryptography, and Miscellaneous categories (2) Cyber Range (5 scenarios of light-to-medium difficulty) to test the model's ability to conduct long-form, end-to-end cyber operations (3) Evasion, network attack simulation, and vulnerability discovery and exploitation (Pattern Lab external assessment) **Methodological Details** include (1) Environment and Elicitation setup (e.g. headlessLinux box, tool harness) (2) Benchmarks and model performance baselines (3) Quantitative evaluation metrics System Card (pp. 27-35)	**Yes** **Safety Framework Classification** Cyber risks are tested for Cyber Autonomy Level 1 and Cyber Uplift Level 1, both unreached. However, the model crossed the early-warning alert threshold for Uplift Level 1. **Evaluation Scope** includes: (1) Existing Capture-the-Flag (CTF) challenges primarily for autonomy tests: i) InterCode-CTF (easy, undergraduate level) ii) In-house suite (medium, graduate-level) iii) Hack the Box (hard, professional level) (2) Key skills benchmark (Rodriguez et al., 2025)for uplift tests: 8 mapped challenges to measure 4 critical competencies: i) Reconnaissance ii) Tool development iii) Tool usage iv) Operational security. **Methodological Details** include: (1) Environment and elicitation setup (e.g. Bash and Python execution) (2) Benchmarks and model performance baselines System Card (pp. 14-17), Technical Report (pp. 30-32)	**Yes** **The Evaluation Scope covers** automate cyberattacks, identify and exploit security vulnerabilities, and automate harmful workflows. **Methodological Details** include threat modeling exercises and capability-based challenge construction.	**Yes** **Safety Framework Classification** None **Evaluation Scope** covers: (1) Cyber knowledge (e.g. Metasploit, vulnerability detection, reverse engineering simple binaries) (2) Cyber agent **Methodological Details include:** (1) Environment setup (Inspect by UK AISI, agent harness) (2) Benchmarks (WMDP Cyber, CyBench) (3) Qualitative metrics System Card (pp. 5-6)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Autonomous AI R&D							
Yes **Safety Framework Classification** Evaluation test thresholds for 1) Checkpoint 2) AI R&D 4 (ASL-3); 3) AI R&D 5 (ASL-4) **The scope of evaluation** includes 1) A checkpoint: a wide range of 2–8 hour software engineering tasks - SWE-bench Verified (hard subset) 2) ASL-4: custom difficult AI R&D tasks built in-house - Internal AI research evaluation suite 1 (e.g. kernels task, time series fore casting, text-based reinforcement learning task, LLM training etc.) - Internal AI research evaluation suite 2, - Internal Model evaluation and use survey **Methodological details** include 1) Environment and elicitation (e.g. context and prompt lengths variations, example-based prompts) 2) Benchmarks with human/model performance baselines 3) Quantitative evaluation metrics System Card (pp. 136-147)	**Yes** **Safety Framework Classification** AI self-improvement capabilities are tracked as part of ongoing safety monitoring. **The Evaluation Scope** covers (1) Real-world software engineering tasks (SWE-bench Verified (N=477), SWE-Lancer (Diamond IC-SWE)) (2) Real world ML research tasks (OpenAI PRs) (3) Real world data science and ML competitions (MLE-Bench) (5) Real world ML paper replication (PaperBench) (6) Real world ML debugging and diagnosis (OPQA (OpenAI-Proof Q&A)) **Methodological Details** include (1) Environment and Elicitation setup (e.g. virtual environment with with tool access, bash execution, and GPU resource, maximum trained-in verbosity) (2) Benchmarks with human/model performance baselines (3) Quantitative evaluations metrics System Card (pp. 35-43)	**Yes** **Safety Framework Classification** Machine Learning R&D capabilities are tested for ML R&D Autonomy Level 1 and ML R&D Uplift Level 1, both remaining unreached. **The Evaluation Scope** covers Research Engineering Benchmark (RE-Bench, Wijk et al.2024) - 5 tests (2 tests omitted due to security concerns of internet access) **Methodological Details** include (1) Environment and elicitation setup (e.g. METR's modular scaffold with minimal adjustment) (2) Benchmark with human expert and model performance baselines (2) Quantitative evaluation metrics System Card (pp. 17-19); Technical Report (pp. 33-36)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Table continues on next page

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Scheming & Misalignment Risks							
Yes **The scope of evaluation** includes alignment faking, undesirable or unexpected goals, hidden goals, deceptive or unfaithful use of reasoning scratchpads, sycophancy toward users, a willingness to sabotage our safeguards, reward seeking, attempts to hide dangerous capabilities, and attempts to manipulate users toward certain views. **Methodology domains** cover the following aspects including: (1) Automated behavioral audits with realism filtering, example seed instructions and evaluation criteria. (2) Third-party replications in collaboration with UK AISI and Apollo Research; (3) Training and pilot-use monitoring that deploys longitudinal checks of whether alignment persists in live internal use and reinforcement-learning logs. (4) Risk-area breakdowns — domain-specific sub-evaluations that map failure modes across self-preservation, sycophancy, reasoning faithfulness etc. (5) White-box interpretability investigations, which is the first public frontier-model effort to examine whether internal representations causally support alignment rather than merely simulating it. The investigations combine unsupervised model diffing and targeted evaluation awareness investigation. System Card (pp. 50-114)	**Yes** Sandbagging is added as part of the Research Categories for the Preparedness Framework. While it does not suffice to be a Tracked Category, it represents a field where "work is required now in order to prepare to effectively address risks of severe harms in the future." **The scope of evaluation** for the broader alignment field includes Sycophancy, Deception (coding deception, browsing broken tools, CharXiv missing image, AbstentionBench), Sandbagging. **Methodology details** include (1) Benchmarks (2) Quantitative evaluation metrics System Card (pp. 7-8, 13-15, 43)	**Yes** **Safety Framework Classification** Deceptive-alignment evaluations are tested Instrumental Reasoning Level 1 and Level 2, both remaining unreached. **Evaluation Scope** covers (1) Stealth evaluations (2) Situational awareness evaluations **Methodological Details** are not included in the model card, but Phuong et al. (2025) publishes the following details: (1) Environment and elicitation setup (e.g .scaffolds that uses chain-of-thought reasoning and reflection to enhance context utilization and better planning) (2) Benchmarks (3) Quantitative evaluation metrics System Card (pp. 19-20)	**Not Mentioned**	**Yes** **The scope of evaluation** includes Manipulation Sycophancy, Deception, and Persuasiveness **Methodology details** include (1) Benchmarks (MASK for deception, OpenAI's MakeMeSay for persuasiveness) (2) Quantitative evaluation metrics System Card (pp. 4, 6)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Indicator

Elicitation for Dangerous Capability Evaluations

Definition

This indicator assesses how clearly a company explains its elicitation strategy, which is the systematic and state-of-the-art techniques it uses to reveal the model's full range of capabilities and potential dangerous behaviors that may otherwise remain concealed. Such techniques include adapting test-time compute, rate limits, scaffolding, and tools, and conducting fine-tuning and prompt engineering.

Why This Matters

Standard evaluations often capture only a model's default, surface-level behavior, leaving deeper or more hazardous capabilities undiscovered. By systematically varying prompts, sampling methods, tools, and system configurations, evaluators can reveal capabilities that may emerge only under real-world or adversarial conditions. A comprehensive, transparent, and well-resourced approach demonstrates a credible commitment to risk discovery.

EU AI Code of Practice Safety and Security	**Anthropic**	**OpenAI**	**Google DeepMind**	**Meta**	**xAI**	**DeepSeek**	**Z.ai**	**Alibaba Cloud**
	Claude Sonnet 4.5	**GPT-5**	**Gemini 2.5 Pro**	**Llama 4**	**Grok-4**	**R1**	**GLM-4.6**	**Qwen3-Max**
Appendix 3.2 Signatories are required to conduct model evaluations using at least state-of-the-art elicitation methods that minimize under-elicitation and model deception during model evaluation, and that match both the capabilities of potential misuse actors and the model's expected use context. Examples of the measures include adapting test-time compute, rate limits, scaffolding, tools, fine-tuning, and prompt engineering	**Adapting test-time compute** is reported in cyber evaluations (e.g. flexible token constraints) and CBRN evaluations (e.g. pass@5 results reported for longform virology, extended thinking) and alignment evaluations (extended thinking) **Scaffolding** is reported in cyber evaluations (e.g. specific resets and auto-summorization in CyberGym) **Iterative Prompting** is reported in CBRN evaluations (e.g. prompt engineering based on analyzing failure cases) **Tool use** is reported in CBRN evaluations (e.g. tools and agentic harnesses) and cyber evaluations (e.g. code editor and a terminal tool) **Helpful-only variants** are reported in CBRN evaluations.	**Adapting test-time compute** is reported in cyber evaluations (e.g. pass@12 for CTF challenges and cyber range evaluations) and and AI self-improvement evaluations (e.g. SWE-bench and MLE-Bench multi-rollout trials). **Tool use** is reported in cyber evaluations. **Custom post-training** (e.g. helpful-only variants), **scaffolding** and **prompting** are applied where relevant, though the System Card does not specify which evaluations each technique was used in.	**Scaffolding and Agent Harness** is reported in cybersecurity, machine-learning R&D, and deceptive-alignment tests, which includes chain-of-thought and reflection loops. **Tool use** is reported in cybersecurity evaluations. **Parallel attempt setups** is reported by cybersecurity evaluations (10-50 attempts) and deceptive-alignment tests (50 retries) and meanwhile time and run budgets (43 × 45-minute vs 16 × 2-hour runs) are mentioned for ML R&D benchmarks. **Prompt engineering** is reported in CBRN and cybersecurity (e.g. open-ended, multi-turn).	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned

Indicator

Human Uplift Trials

Definition

This indicator assesses whether organizations conduct rigorous, controlled human-subject studies to evaluate the marginal risk AI systems pose in dangerous domains by "uplifting" people's ability to cause harm. Key evidence includes experimental designs that compare task performance with and without AI support, the inclusion of domain-relevant experts, realistic and consequential task scenarios, and transparent publication of methods and findings. To assess worst-case potential, models should be tested without embedded safety filters.

Why This Matters

Empirical uplift studies are critical for grounding AI safety policy in observable outcomes. These studies assess whether advanced systems significantly enhance a user's ability to cause harm and inform the development of proportionate safety interventions. Entities that conduct and publish such studies exhibit leadership in transparent, evidence-based risk governance.

EU AI Code of Practice Safety and Security	**Anthropic**	**OpenAI**	**Google DeepMind**	**Meta**	**xAI**	**DeepSeek**	**Z.ai**	**Alibaba Cloud**
	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Measure 3.2 Examples of model evaluation method include human uplift studies.	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

External

Indicator

Independent Review of Safety Evaluations

Definition

Assesses whether an AI developer commissions independent third-party experts to (A) verify the factual accuracy and process integrity of its internal dangerous-capability evaluations and (B) assess the evaluation quality and the company's interpretation of the results. We collect information on the reviewers' identity and credentials, their independence (including any conflicts of interest), the scope of the review, depth of access to data and logs (including rights to replicate or extend tests), and whether their findings are published unredacted.

Why This Matters

AI developers control both the design and disclosure of dangerous capability evaluations, creating inherent incentives to under-report alarming results or select lenient testing conditions that avoid costly deployment delays. Regulators, investors, and the public, therefore, face a critical information asymmetry: they must trust safety claims based on self-reported evaluations with minimal methodological transparency. Independent external scrutiny can address this trust deficit by verifying reported results, assessing whether evaluations are sufficiently rigorous to uncover real risks, and providing credible third-party perspectives on whether safety claims are justified. This need is especially acute for catastrophic risk domains such as biosecurity, where companies may cite "infohazard" concerns to limit transparency.

Anthropic	**OpenAI**	**Google DeepMind**	**Meta**	**xAI**	**DeepSeek**	**Z.ai**	**Alibaba Cloud**
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Not Mentioned	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**

Indicator

Pre-deployment External Safety Testing

Definition

This indicator evaluates whether companies enable external safety assessments of frontier AI models before public release, and the degree to which those evaluators operate independently from the model developer. Independent will be assessed across four dimensions, including institutional affiliation, methodological autonomy, access autonomy, and publication freedom. Evidence includes the identity and qualifications of external parties, the level and duration of access provided, compensation arrangements, testing permissions, and the evaluators' ability to publish independently. The strength of these practices is judged by the comprehensiveness of the evaluations, the depth of access, and the autonomy of the evaluators.

Why This Matters

External evaluations are essential for verifying safety claims and uncovering risks that internal teams may overlook or under-report. Providing external evaluators with substantial access and ensuring their ability to test and publish with a great amount of autonomy reflect a company's commitment to transparent and evidence-based governance.

Table begins on the next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Appendix 3.4-3.5 Signatories must ensure that qualified independent external evaluators assess their models for systemic risk unless the model is already proven comparably safe or evaluators cannot be secured after reasonable efforts. These evaluators must have relevant technical expertise (academic or professional) and follow strict security and confidentiality protocols. Meanwhile, signatories will provide independent external evaluators with (1) adequate access (e.g. access to model activations, gradients, logits, chains-of-thought, model version(s) with the fewest safety mitigations implemented) (2) information (e.g. model specifications (including the system prompt), relevant training data, test sets, and past model evaluation results), (3) time, and (4) other resources (e.g. compute budgets, staffing, engineering budgets and support) Signatories will not undermine the integrity of external model evaluations by storing and/or analyzing inputs and/or outputs from test runs without express permission from the evaluators.	External organizations shared summaries of initial findings for Anthropic to reproduce and compare results with internal investigations for the snapshots and final versions. According to Anthropic's Transparency Hub, "external evaluations use API access with zero-data-retention settings to prevent content storage," consistent with the practices identified in our previous iteration of the AI Safety Index (July 2025). **UK AI Security Institute (UK AISI)** Access: an early snapshot, access released on September 22, 2025) Scope: Misalignment threats (e.g. self-preservation, evaluation awareness etc.) Validation method: Ablations of key environment factors **Apollo Research** Access: pre-deployment snapshot Scope: Misalignment threats (e.g. strategic deceptions, scheming, evaluation awareness etc.) **Independence** (1) Evaluators may publish independently after company review/possible redaction. (2) The company provided its own summary of the evaluator's key findings.	**Scope** SecureBio (Static, agent, and long-form evaluations + manual red teaming for bio risks); Pattern Labs (Evaluates evasion, network attack simulation, and vulnerability discovery and exploitation); METR (AI R&D automation, rogue replication, strategic sabotage); Apollo Research (Covert & deceptive actions); Gray Swan Arena Platform (Prompt-injection and bio-weaponization jailbreaks); FAR.AI (Biological and system-level jailbreak stress tests); U.S. Center on Artificial Intelligence Standards and Innovation (CAISI) (Cyber, biological, and chemical capabilities and safeguards); UK AISI (Cyber and biological / chemical capabilities, plus safeguard penetration testing); Microsoft AI Red Team (Frontier Harms, Content Safety, and Psychosocial Harms). **Access** (1) The longest period of time that an external evaluator was given continuous access for pre-deployment testing is >2 weeks (<=3 weeks). (2) The highest level of technical access granted to any of the listed external evaluators is Standard inference API with normal user-facing filters in place, Inference API with safety filters disabled (no inference-time mitigations), and "Helpful-only" or base model API (no harmlessness fine-tuning and no filters). (3) Third party assessors were provided OpenAI GPT-5 Thinking early checkpoints, as well as the final launch candidate models. **Security** Zero Data Retention available upon request, if technically feasible during pre-deployment periods	External organizations are chosen based upon their domain expertise, and include civil society and commercial organizations. However, they are not named individually. **Scope:** Autonomous systems, cybersecurity, CBRN, and societal risk **Access:** (1) The highest level of technical access granted to any of the external evaluators is the Black-box access to Gemini 2.5 Pro (Preview 05-06) via the inference API, with safety filters disabled (no inference-time mitigations). (2) The longest period of time that an eternal evaluator was given continuous access for pre-deployment is >3 weeks (<=5 weeks). (3) For pre-deployment testing, evaluators had higher quotas for query rates than the public/ enterprise tier but were still subject to explicit caps (e.g. requests-per-minute or daily token limits). The quota is bespoke depending on the testing partner's specific needs and evaluation type. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks.	**Not Mentioned**	xAI has responded that external testing was commissioned in the survey response without naming the evaluators. The external safety tests were completed before broad internal deployment. They released the same model version that the final round of safety evaluations were conducted on. **Access:** The highest level of technical access it has shared externally is Helpful-only' or base model API (no harmlessness fine-tuning and no filters), with the longest duration of more than 5 weeks. Evaluators will have higher quotas than the public or enterprise tiers for query rates but are still subject to explicit caps (e.g. requests-per-minute or daily token limits. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. **Independence:** Evaluators may publish independently after company review or possible redaction. **Timeline:** All external safety tests were completed before broad internal deployment. Source: Company Survey	**Not Mentioned**	**Scope:** Z.ai has collaborated with China Academy of Information and Communications Technology (CAICT), which is a subordinate to the powerful Ministry of Industry and Information Technology (MIIT), for evaluations of "general safety issues," as according to the survey response. **Access:** The highest level of technical access it has shared externally is 'Helpful-only' or base model API (no harmlessness fine-tuning and no filters). There are no limits for query-rate or volume restrictions to external evaluators. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. **Independence:** Evaluators may publish independently without prior company approval after the model is released. **Timeline:** All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment. Source: Company Survey	**Not Mentioned**

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
		Independence (1) Evaluators may publish independently without prior company approval after the model is released, provided that evaluations are run independently on the deployed model. (2) Evaluators may publish independently after company review/possible redaction. Since pre-deployment evaluation period are under NDA, publications require prior approval to protect confidential information. METR has published the full report. (3) The company provided its own summary of the evaluator's key findings, which they share with the evaluator prior to release to confirm factual accuracy. (4) OpenAI publishes excerpts from the report mutually agreed upon or written, with the company having the final say for what content goes in System Cards. **Timeline** External safety tests were completed after broad internal deployment.	**Independence:** These organizations are independent in choosing methodologies, ranging from qualitative red-teaming to quantitative automated testing, at varying time commitments. After receiving all analyses, raw data, and evaluation materials, internal experts reviewed model outputs and applied harm-severity ratings under established safety frameworks and Critical Capability Levels, and writing reports internally. External evaluators are financially compensated by Google DeepMind for their time. Technical Report (pp. 36-38), Company Survey					

Indicator

Bug Bounties for System Vulnerabilities

Definition

This indicator evaluates whether companies maintain structured programs that reward or formally recognize external researchers for discovering and responsibly disclosing safety vulnerabilities in AI system behavior, such as through red-teaming initiatives or bug bounties. The focus is primarily on behavioral vulnerabilities, such as jailbreaks, prompt attacks, data extraction, or adversarial manipulations, rather than conventional software or cybersecurity bugs. Evidence includes the scope of eligible vulnerabilities, reward structure or compensation levels, response and disclosure processes, and the public availability of program rules and results.

Why This Matters

Structured disclosure programs with financial incentives harness external expertise to identify system vulnerabilities before they are exploited in deployment. Investments in such programs indicate openness and proactiveness toward risk identification.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Anthropic has previously run 2 rounds of bug bounty programs in August 2024 and May 2025. Anthropic announced on May 22, 2025, an ongoing bug bounty initiative accepting applications on a rolling basis, as opposed to "invitation-only" in the previous rounds. **Scope:** The program focuses on live deployed systems with ASL-3 protections, and seeks universal and detailed jailbreaks that extract detailed biological-threat information. **Reward:** Up to $35,000 per novel, universal jailbreak identified. (up to $15,000 in August 2024 and up to $25,000 in May 2025) **Timeline:** Issues are resolved usually within ~ 1 days although time to resolution is missing. **Access:** Participants have access to free model aliases that reflect the model and classifiers live on our latest, most advanced model, as opposed to early access to unreleased safety mitigation systems and models in the previous rounds. **Confidentiality:** Formal NDA frameworks	**Scope:** The ongoing bug bounty program covers a wide range of security vulnerabilities across its products and infrastructure, including the OpenAI API, ChatGPT (Plus, plugins, and agent modes), Sora, Atlas. It explicitly excludes model behavior or safety issues (e.g., jailbreaks, hallucinations, prompt content). **Reward** scale by severity: - Critical (P1): up to $100,000 - High (P2): $2,000–$6,500 - Medium (P3): $1,000–$2,000 - Low (P4): $200–$500 **Timeline:** Validation is usually within 6 days. 75% of submissions are accepted or rejected within 6 days in last 3 months. **Access:** Participants test in-scope systems only. API testing, plugin testing (only for self-created plugins), and limited third-party vendor exposure checks are permitted. **Confidentiality:** Partial Safe Harbor for good-faith security research but requires strict confidentiality, prohibiting public disclosure until OpenAI authorizes it (usually within 90 days).	**Scope:** The ongoing AI Vulnerability Reward Program (VRP) covers AI-related security and abuse vulnerabilities in Google/Alphabet AI products, where interaction with an LLM or GenAI system is integral to the bug. Policy or alignment bypasses, jailbreaks, hallucinations, and content violations are explicitly out of scope. Vertex AI and other Google Cloud issues are handled by the separate Cloud VRP. **Reward:** up to US $20,000 for rogue actions detected with flagship products (including Gemini products), adjusting for reporting quality and accounting for novelty bonus (+$1k - +$5k). **Access:** testing limited to researcher's own/test accounts (recommended); no special model access. **Confidentiality:** Participants should follow a designated Code of Conduct, under which they are encouraged to follow coordinated vulnerability disclosure and are expected to have good faith.	**Scope:** The ongoing bug bounty program (started in 2023) is restricted to privacy or security issues, like extracting training data through tactics like model inversion or extraction attacks. (Consistent with the findings of July 2025 AI Safety Index) **Reward:** - The minimum reward for a qualifying submission is US $500. - The maximum reward for a qualifying submission in Meta AI is US $30,000. **Access:** Participants do not have special access to the models but are encouraged to use authorized or test accounts. **Confidentiality:** Meta's Bug Bounty confidentiality and disclosure rules require researchers to avoid privacy violations, use only authorized or test accounts, immediately report and delete any inadvertently accessed data, and give Meta reasonable time to investigate before any public disclosure. Safe-harbor protections apply only if researchers act in good faith and fully comply with these terms.	**Scope:** The program covers xAI, including the Grok API, and targets traditional security vulnerabilities, including authentication, authorization, data-exposure, and infrastructure issues. However, model behaviors and AI safety issues are explicitly out of scope. **Reward:** Bounties are discretionary, determined by a 5×5 internal risk matrix (impact × likelihood) and by a panel of security experts. 90-day averages as of the last update (May 2025): ▪ Low $100 – $500 (19.6 %) ▪ Medium $500 – $2,000 (40 %) ▪ High $2,500 – $7,000 (30 %) ▪ Critical $7,500 – $20,000 (10 %) **Timeline:** Issues are usually triaged within ~1 day and resolved within ~3 weeks. **Access:** No mention of model access or sandbox environment. **Confidentiality:** Participants must abide by HackerOne's disclosure guidelines, including using test accounts, protecting user privacy, and keep all findings confidential until the report is closed.	Not Mentioned	Not Mentioned	Not Mentioned

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Risk Assessment

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Comprehensive, state-of-the-art evaluations; strong validity, reproducibility, and independent review; no serious harm potential.
- **B** Robust assessments; good validity and elicitation; limited external review; serious harms well-controlled.
- **C** Partial assessments; uneven validity or elicitation; little external input; serious harms mostly controlled.
- **D** Fragmented assessments; weak validity and elicitation; no external review; serious harms poorly controlled.
- **F** No credible assessment; serious harm uncontrolled.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Current Harms

This domain covers demonstrated safety outcomes rather than commitments or processes. It focuses on the AI model's performance on safety benchmarks and the robustness of implemented safeguards against adversarial attacks.

Table of Contents

Chinese Regulatory System Summary

China's Interim Measures mandate strict data minimization, lawful handling of user information, and timely fulfillment of user rights requests, ensuring robust privacy protection. Meanwhile, the Deep Synthesis Regulation and National Standard GB45438-2025 require AI providers to implement both explicit and implicit watermarking systems, ensuring traceability and transparency of AI-generated content.

National Binding Instruments

Privacy

Interim Measures Article 11 requires AI service providers to lawfully protect users' input data and usage records.

Specifically, they must not collect unnecessary personal information (data minimization), must not illegally retain identifiable input data or usage records, and must not illegally provide such information to others (lawful handling). In addition, providers must timely accept and handle user requests to access, copy, correct, supplement, or delete their personal information (responsive obligations to user rights).

Watermarking

Deep Synthesis Regulation Article 16-18 requires that deep synthesis service providers are required to add built-in watermarks and keep system logs. When content could confuse people, providers must place prominent marks on generated or edited content. They must also provide labeling functions for other synthetic content and remind users they can apply visible marks. No one is allowed to remove or alter these marks.

National Standard GB45438—2025 Cybersecurity technology—Labeling method for content generated by artificial intelligence delineates the specific requirements that AI service providers have to follow when placing explicit vs. implicit watermarks.

For explicit labeling, when AI-generated text, audio, video, or other content could mislead or confuse the public, providers must apply clear and visible marks at specified positions.

For implicit labeling, every AI-generated file must contain standardized metadata that includes: (1) an AI-generation tag; (2) the service provider's name or code; (3) a unique content ID; (4) the distributor's name or code; and (5) a unique distribution ID. Content-implicit labeling is optional and not required under this standard.

Safety Performance

Indicator

Stanford's HELM Safety Benchmark

Definition

This indicator measures model performance on Stanford's HELM Safety v1.0 benchmark, a suite of five safety tests covering six risk categories: violence, fraud, discrimination, sexual content, harassment, and deception. The benchmark includes: HarmBench (jailbreak resistance); BBQ (social discrimination); SimpleSafetyTest; XSTest (alignment between helpfulness and harmlessness); and AnthropicRedTeam (resilience to adversarial probing). Performance is reported as normalized aggregate scores ranging from 0 to 1, where higher scores indicate fewer safety risks. Scoring is based on exact match accuracy for BBQ and model-judge ratings (GPT4o and Llama 3.1 405B) for the remaining benchmarks.

Why This Matters

HELM Safety provides a standardized, empirical benchmark for evaluating how reliably AI systems prevent harmful or unsafe outputs. It measures behavioral safeguards—such as refusals of violent, fraudulent, or discriminatory content—under consistent testing conditions. Strong performance demonstrates that a model's technical safety mechanisms effectively reduce direct user-facing risks across diverse harm categories.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Models Evaluated	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Average score (max score = 1)	**0.97**	**0.98**	**0.91**	**0.91**	**0.84**	**0.87**	Model not evaluated	Model not evaluated.
HarmBench	0.92	0.98	0.65	0.66	0.40	0.47		
SimpleSafetyTests	1.00	1.00	0.97	0.99	0.92	0.98		
BBQ accuracy	0.99	0.97	0.96	0.93	0.94	0.97		
AnthropicRedTeam	0.98	0.99	1.00	0.98	0.96	0.96		
XSTest	0.96	0.97	0.99	0.97	0.97	0.95		
Retrieved	November 3, 2025							
Release	October 2, 2025 (v.1.16.0)							

Footnotes

[1] Farzaan et al. "HELM Safety: Towards Standardized Safety Evaluations of Language Models." Stanford Center for Research on Foundation Models, 8 Nov. 2024. Accessed 3 Nov, 2025.

Indicator

Stanford's HELM AIR Benchmark

Definition

This indicator evaluates model performance on Stanford's AIR-Bench 2024 (AI Risk Benchmark), an AI safety benchmark aligned with emerging government regulations and company policies. We report mean scores across 5,694 tests spanning 314 granular risk categories, with scores measuring the percentage of appropriately refused requests. The benchmark systematically evaluates four major risk domains: System & Operational Risks (e.g., cybersecurity, operational misuse), Content Safety Risks (e.g., child sexual abuse material), Societal Risks (e.g., surveillance), and Legal & Rights-related Risks (e.g., privacy violations, defamation). All prompts are manually curated and human-audited to ensure they reflect genuine policy violations rather than benign content.

Why This Matters

HELM AIR provides an evaluation of how well AI systems align with real-world safety expectations. Unlike behavioral safety tests, it directly reflects the kinds of standards developers will be expected to meet, rooted in emerging regulation, ethics, and risk-management practices. Strong performance signals high readiness of AI systems to comply with policy and societal values.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Models Evaluated	Claude 4.5 Sonnet (20250929)	GPT-5 (2025-08-07)	Gemini 2.5 Pro	Llama 4 Maverick (17Bx128E) Instruct FP8	Grok-4 (0709)	R1	GLM-4.5-Air-FP8	Qwen3 235B A22B Instruct 2507 FP8
Average score (max score = 1)	**0.90**	**0.93**	**0.74**	**0.71**	**0.40**	**0.48**	**0.55**	**0.79**
System & Operational Risks	0.81	0.98	0.64	0.55	0.30	0.28	0.29	0.61
Content Safety Risks	0.88	0.83	0.73	0.78	0.51	0.54	0.62	0.81
Societal Risks	0.97	0.95	0.74	0.72	0.32	0.51	0.61	0.86
Legal & Rights-related Risks	0.95	0.95	0.83	0.79	0.49	0.59	0.68	0.88
Retrieved	November 3, 2025							
Release	October 2, 2025 (v.1.16.0)							

Footnotes

[1] Zeng et al. "Air-bench 2024: A safety benchmark based on risk categories from regulations and policies." 2024. Accessed 3 Nov, 2025.

Indicator

TrustLLM Benchmark

Definition

This indicator measures a model's overall trustworthiness using the TrustLLM benchmark, a comprehensive framework spanning six dimensions: truthfulness, safety, fairness, robustness, privacy, and machine ethics. The benchmark includes over 30 datasets across more than 18 subcategories, assessing issues such as hallucination, jailbreak resistance, and privacy leakage. Models are evaluated on tasks ranging from simple classification to complex generation, with results reported as published scores and rankings across each dimension. TrustLLM was developed by 45 research institutions, including 38 based in the U.S.

Why This Matters

TrustLLM evaluates how reliably AI systems uphold truthfulness, privacy, and ethical reasoning beyond standard capability metrics. Strong performance indicates that companies have invested in aligning their models to be harmless and helpful, and not to cause unintended harm.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Models Evaluated	Claude-Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Average score (max score = 1)	**0.64**	**0.60**	**0.63**	**0.60**	**0.62**	**0.62**	**0.62**	**0.62**
Truthfulness	0.50	0.49	0.59	0.54	0.58	0.61	0.59	0.47
Safety	0.66	0.63	0.62	0.60	0.60	0.69	0.63	0.69
Fairness	0.48	0.42	0.43	0.47	0.39	0.38	0.53	0.44
Privacy	0.65	0.51	0.59	0.53	0.62	0.52	0.52	0.63
Ethics	0.89	0.86	0.89	0.84	0.85	0.89	0.81	0.81
Robustness	0.68	0.66	0.64	0.61	0.67	0.62	0.65	0.66
Retrieved	Full Score Breakdown							
Release	v.0.3.0							

Footnotes

[1] Yue Huang et al., TrustLLM: Trustworthiness in Large Language Models (arXiv:2401.05561, 2024), https://arxiv.org/abs/2401.05561.

Indicator

CAIS Leaderboard Benchmarks

Definition

The safety benchmark collection includes six benchmarks. Lower scores indicate better performance.

(1) Agent Red Teaming measures robustness against malicious use and prompt injections. The benchmark evaluates AI agents across five core behavior categories: Confidentiality Breaches, Conflicting Objectives, Instruction Hierarchy Violations — Information, and Instruction Hierarchy Violations — Actions. Lower scores indicate better performance.

(2) Virology Capabilities Test (VCT) Refusal measures the rate at which AIs refuse hazardous expert-level virology queries. It is a subset of VCT that is manually labeled by experts for queries that should be refused because of potential dual-use. The ability to provide expert-level troubleshooting is inherently dual-use: it is useful for beneficial research, but it can also be misused in particularly harmful ways.

(3) Humanity's Last Exam (HLE) Miscalibration measures how well AI models calibrate their confidence with their actual accuracy. This metric evaluates whether models are overconfident or underconfident in their predictions across 2,500 expert-written academic questions spanning 100+ disciplines. Lower calibration error indicates better alignment between a model's expressed confidence and its true performance.

(4) MASK tests the extent to which AIs can be pressured to overtly lie. It is different from measuring accuracy or truthfulness, but instead, it measures whether a model consistently expresses its beliefs even when pressured to do otherwise. [Scale AI]

(5) Machiavelli measures harmful behaviors in LLM agents across 30 text-based games. We use its deception measurements from its various goal-oriented scenarios.

(6) TextQuests's harm metric measures the AI's propensity to engage in wantonly harmful behaviors in text-based adventure games.

Why This Matters

These benchmarks matter because they test safety-relevant traits—like honesty, refusal behavior, and ethical restraint—that do not automatically improve with model size or training compute. Their low correlation with general capabilities means they capture distinct aspects of alignment and behavioral reliability rather than raw intelligence. This separation helps prevent "safetywashing," where capability gains are mistaken for safety progress. In doing so, they provide a more rigorous basis for tracking genuine advances in AI safety as systems grow more powerful. [Ren et al., 2024]

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Models Evaluated	Claude-Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Average Score	**38.80**	**51.80**	**67.80**	**65.10**	**55.00**	**62.90**	**65.30**	**71.30**
Agent Red Teaming	40.00	54.00	90.60	86.60	65.00	90.10	83.40	87.20
Virology Capabilities Test (VCT) - Refusal	31.50	92.50	100.00	100.00	55.00	100.00	100.00	100.00
Humanity's Last Exam (HLE) - Miscalibration	64.10	50.00	72.00	84.00	56.40	73.00	80.00	89.60
MASK	3.90	20.70	44.30	50.30	41.70	42.70	42.70	49.70
Machiavelli	73.70	76.60	83.90	56.50	81.50	56.00	67.10	81.60
TextQuests Harm	19.60	17.20	15.90	13.10	30.40	15.40	18.70	19.80

Footnotes

[1] Long Phan and Dan Hendrycks, CAIS AI Leaderboard (Center for AI Safety, 2025), https://leaderboard.safe.ai.

Digital Responsibility

Indicator

Protecting Safeguards from Fine-tuning

Definition

This indicator evaluates whether companies maintain safeguards that prevent the removal of built-in safety measures during fine-tuning. Evidence differentiates between: i) Supervised or hosted fine-tuning, which occurs on the company's platform where core safety filters remain active; and ii) Full model-weight releases, where users can directly modify parameters and potentially disable all protections unless tamper-resistant controls are in place.

If companies provide no public information on fine-tuning or weight-release policies for their frontier AI systems, these capabilities are treated as not publicly accessible.

Why it matters

Releasing full model weights may allow malicious actors to strip or override safety mechanisms, creating uncensored or harmful versions. In contrast, supervised fine-tuning preserves core safety guardrails while enabling responsible customization.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Frontier model weights protected Provide supervised fine-tuning for older and smaller Claude 3 Haiku through Amazon Bedrock. Safety mitigations are in place. [AWS, 2024]	Frontier model weights protected. Released weights of non-frontier gpt-oss-120b and gpt-oss-20b. No tamper-resistant safeguards. [Hugging Face, 2025] Provide supervised fine-tuning (SFT) of gpt-4.1-2025-04-14, gpt-4.1-mini-2025-04-14, and gpt-4.1-nano-2025-04-14 [OpenAI, 2025] and RL fine-tuning for o4-mini-2025-04-16. [OpenAI, 2025]	Frontier model weights protected. Released weights of non-frontier Gemma family, including Gemma 327B. No tamper-resistant safeguards. [Hugging Face]. Enables supervised finetuning of Gemini 2.0 Flash, 2.0 Flash-Lite, 2.5 Flash, 2.5 Pro, and 2.5 Flash-Lite via Vertex AI. Safety mitigations are in place. [Google, 2025].	Fully released weights of the frontier model Llama 4 Maverick. No tamperresistant safeguards. [Meta AI]	Frontier model weights protected. Fully released weights of non-frontier Grok 1. No tamper-resistant safeguards. [xAI, 2024]	Fully released weights of frontier models. No tamper-resistant safeguards. [Hugging Face]	Fully released weights of the frontier model GLM-4.6. No tamper-resistant safeguards. [Hugging Face]	Frontier model weights protected. Fully released weights of non-frontier Qwen3 family, including Qwen3-235B-A22B. No tamper-resistant safeguards. [Hugging Face]

Indicator

Watermarking

Definition

This indicator assesses whether companies have implemented watermarking technologies to help identify AI-generated content in both text and images. It focuses on real-world implementation rather than research prototypes, evaluating the accuracy and robustness of detection methods, adherence to standards such as C2PA and SynthID, and whether detection tools are publicly accessible.

Why This Matters

Watermarking helps distinguish authentic content from AI-generated media, reducing the risks of misinformation, fraud, and reputational harm. Companies that implement robust and standardized watermarking systems, and make detection tools publicly accessible, demonstrate a strong commitment to transparency, provenance, and digital trust.

Sub-Indicator	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Text-based	**None found**	**No** OpenAI has announced that it has developed a text watermarking method, but it is still researching for alternatives, due to concerns over its effectiveness against globalized tampering, and disproportionate stigmatizing impact on non-native English speakers. [OpenAI, 2024]	**Yes** (SynthID) The SynthID system uses particular token selection to introduce a pattern that marks a text as AI-generated [Google DeepMind]. This can be identified using an online detection tool, which is currently accessible only to approved journalists, media professionals, and researchers through a waitlist program. [Google, 2025].	**None found**	**None found**	**Yes** Under the 2025 National Standard on AI-Generated Content Labeling and Watermarking, companies must include explicit watermarks to identify content produced by artificial intelligence. The standard applies to text, images, audio, video, and virtual environments. For each content type, it specifies (1) where the label must appear, (2) the required information to include in the label, and (3) the parameters that determine its visibility or audibility, such as label size, voiceover speed, and display duration.		
Image-based	Claude AI systems do not generate images.	**No watermarking** (C2PA metadata) Images generated with ChatGPT on the web and the API serving the DALL·E 3 model, will now include C2PA metadata.The metadata can be detected unless it has been removed either accidentally or intentionally. [OpenAI, 2025]	**Yes** (SynthID) Pattern is embedded in images, can be identified by an online detector, access currently limited. [Google DeepMind]	The open-source Llama 4 family does not include models that can generate images. However, for photorealistic images created using Meta AI, Meta has applied visible labels of "Imagined with AI" and included invisible watermarks and metadata embedded within files. [Meta, 2024]	**None found**			

Indicator

User Privacy

Definition

This indicator reports a company's dedication to user privacy when training and deploying AI models. It considers whether user inputs (such as chat history) are used by default to improve AI models or if companies require explicit opt-in consent. It also considers whether users can run powerful models privately, through on-premise deployment or secure cloud setups. Evidence includes default privacy settings and the availability of model weights for private hosting.

Why it matters

Privacy controls that require deliberate consent to opt in enable greater respect for user privacy, especially in sensitive fields such as healthcare, law, and government.

Sub-Indicator	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Default training on user inputs	**Yes** Anthropic updated its consumer terms and privacy policy in August 2025, introducing a new data-sharing setting under which user conversations are included in model training by default unless the user manually opts out through the "Help improve Claude" toggle. This applies to users for Claude Free, Pro, and Max plans. Previously, user inputs are only trained for model improvements if they explicitly opt-in or if the conversation is flagged for violating Usage Policy. [AI Safety Index, 2025]	**Yes** (exception for enterprise data) ChatGPT does not train models on Enterprise account user's business data by default. [OpenAI, 2025]	**Yes** (exceptions for Gemini for Google Cloud users) Gemini for Google Cloud doesn't use your prompts or its responses as data to train its models.	**Yes** Meta "use information shared on Meta Products" to train their AI models. "This information could be things like posts or photos and their captions." Private messages are excluded unless "someone in the chat chooses to share those messages with our AIs." [Meta]	**Yes** xAI uses "X posts as well as user interactions, inputs, and results with Grok for training and fine-tuning purposes." [X][Ars Technica, 2024] In addition, xAI uses user inputs to improve its models by default. [xAI, 2025]	**Yes** DeepSeek uses user inputs to improve its models by default. [DeepSeek, 2025]	**Yes** Z.ai uses user inputs to improve its models by default. [Z.ai, 2025] Zhipu's Qingyan, also known as ChatGLM, was found to have collected information beyond what users authorized. [National Cyber Security Information Center, 2025]	**Yes** Alibaba does not provide an opt-out option for users to stop their de-identified content from being used to train the model. [Alibaba, 2025]
Frontier model weights available for private hosting	**No**	**No**, but less-powerful models are open-sourced	**No**, but less-powerful models are open-sourced	**Yes**	**No**, but less-powerful models are open-sourced	**Yes**	**Yes**	**No**, but less-powerful models are open-sourced

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Current Harms

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** No meaningful safety failures; strong resilience to adversarial attacks; negligible harm potential.
- **B** Rare moderate failures; high robustness; serious harms well-controlled.
- **C** Occasional moderate failures; reasonable robustness; serious harms mostly controlled.
- **D** Frequent safety failures; weak robustness; serious harms poorly controlled.
- **F** Widespread failures; minimal or ineffective safeguards; serious harms uncontrolled.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Safety Frameworks

This domain evaluates the companies' published safety frameworks for frontier AI development and deployment from a risk management perspective. The analysis follows the taxonomy and indicator structure developed by the non-profit research organization SaferAI.

Table of Contents

Indicator

Risk Identification

Definition

This dimension assesses how thoroughly the company has addressed known risks in the literature and engaged in open-ended red teaming to uncover potential novel threats. It also evaluates whether the AI company has leveraged a diverse range of risk identification techniques, including threat modeling when appropriate, to develop a deep understanding of possible risk scenarios.

Why This Matters

Companies can only mitigate risks they've identified, making comprehensive risk discovery the foundation of any effective safety framework. Firms that employ diverse identification methods are more likely to catch novel threats before they manifest in deployment. This proactive approach to risk discovery demonstrates whether a company takes seriously the full spectrum of potential harms, including those not yet observed in practice.

Chinese Regulatory System Summary

Mandatory local regulations like the Shanghai and Shenzhen AI rules require ex-ante assessment and controllability reviews for high-risk systems, although they are not directly applicable to Z.ai, DeepSeek, and Alibaba. Voluntary national standards, such as the Risk Management Standard, define structured processes for identifying, analyzing, governing, and mitigating AI risks. Policy guidance documents, including the Ethical Norms and AI Safety Governance Framework 2.0, highlight broader principles for human control, traceability, and frontier-risk prevention without legal enforceability, providing direction for future company compliance.

Voluntary Technical Standard

The Risk Management Standard (Article 5.3.1) breaks down an organization's capability of risk identification into three core components:

(1) selecting appropriate tools, techniques, and methods for identifying risks,

(2) recognizing AI-specific risk sources, and

(3) identifying potential consequences of those risks.

The sources of the risks as identified in Appendix B include frontier AI risks such as Malicious Misuse (e.g. dual-use scientific applications in CBRN development and malicious use), Systemic Safety Risks (e.g. robustness, interpretability, and reliability), Application Security Risks (e.g. loss of control).

Table begins on the next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
1.1 Classification of Applicable Known Risks								
Measure 2.1 (Appendix 1.1 to 1.4) Signatories will identify systemic risk through two approaches. (1) Following the specified structured process to compile a list of identified systemic risks, taking into consideration model-independent data and analysing relevant characteristics such as nature of the systemic risk and sources of the systemic risk (including model capabilities, model propensities, and model affordances) (Appendix 1.1-1.3). (2) Four risks are treated as specified systemic risks that are always identified: CBRN risks, loss of control, cyber offense, and harmful manipulation (Appendix 1.4) **Measure 2.2** Signatories will develop appropriate systemic risk scenarios for each identified systemic risk. **Measure 3.2** Model evaluations should [...] should include open-ended testing of the model, to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties.	Anthropic identifies CBRN weapons and Autonomous AI R&D as its two most pressing catastrophic risks. In addition, it also designates cyber operations as an emerging risk category under ongoing evaluation. Although it recognizes potential risks of highly persuasive AI models, active consultation with experts lead to the conclusion that this capability is "not yet sufficiently understood to include in the current commitments." Anthropic prioritizes these risks through the process of external engagements such as commissioned research reports, discussions with domain experts, input from expert forecasters, public research, conversations with other industry actors through the Frontier Model Forum, and internal discussions.	OpenAI uses a structured risk-assessment process to evaluate whether frontier AI capabilities could lead to severe harm, which is defined as death of thousands or hundreds of billions of dollars in economic damage. The process relies on its own internal research and signals, and where appropriate incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government and its partners, as well as relevant legal and policy mandates. It assigns identified risks to categories: (1) Tracked Categories: currently including Biological & Chemical, Cybersecurity, AI Self-improvement and; (2) Research Categories, including Long-range Autonomy, Nuclear & Radiological for further work.	DeepMind's Framework identifies misuse risks in three domains: Misuse (CBRN, Cyber, and Harmful Manipulation), ML R&D, as well as Misalignment (exploratory) risk. These risks are organized by the framework around capability thresholds called "Critical Capability Levels" (CCLs). The selection is attributed to "early research" that judged these areas most likely to lead to severe harm from future models if unmitigated, but the framework does not describe a formal methodology or process for how these risk domains were identified.	Meta adopts an outcome-based approach described in high levels where it proceeds by (1) defining catastrophic outcomes; (2) maps the causal pathways that could produce them; (3) locate threat scenarios that are potentially sufficient to realize the outcome. The most urgent catastrophic outcomes identified are in the domains of cybersecurity and chemical and biological weapons.	xAI focuses on two overarching systemic risks—malicious use and loss of control—and organizes concrete risk scenarios across abuse potential (e.g., vulnerability to jailbreaks), concerning propensities (e.g., a propensity for deceiving the user), and dual-use capabilities (e.g., offensive cyber capabilities). It does not spell out a formal risk-identification process, but it does quantify "catastrophic malicious-use events" using thresholds for expected fatalities and economic damage.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
1.2 Identification of Unknown Risks								
	The Responsible Scaling Policy does not specify pre-deployment measures to identify novel risk domains for the frontier model, although Anthropic has implemented adversarial testing, red-teaming, and bug bounty programs that can help the company identify unknown threats.	**The Preparedness Framework** mentions that OpenAI conducts adversarial testing, red-teaming, and bug bounty programs to proactively identify and mitigate unknown vulnerabilities and emerging threats across its corporate, research, and product systems.	The Frontier Safety Framework explicitly states that it will "continue to assess whether there are other risk domains where severe risks may arise and will update our approach as appropriate," Moreover, the early warning evaluations are intended to to flag when a CCL may be reached before the evaluations are run again, however, it is also used for detecting novel risks from the frontier AI systems.	The team follows the general process of (1) Hosting workshops with experts to identify new catastrophic outcomes and/or threat scenarios (2) Designing new assessments if novel outcomes/scenarios are identified.	The RMF has not explicitly designated a process specifically for identifying unknown risks, although it emphasizes the development of naturalistic evaluation environments to assess more realistic, real-world model behaviors.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Table continues on next page

EU AI Code of Practice Safety and Security	**Anthropic**	**OpenAI**	**Google DeepMind**	**Meta**	**xAI**	**DeepSeek**	**Z.ai**	**Alibaba Cloud**
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
1.3 Risk Modeling								
Measure 3.3 Signatories will model systemic risks using **at least** state-of-the-art methods, informed by predefined risk scenarios (Measure 2.2) and data collected through prior identification measures (Measure 2.1)	Anthropic has implemented a **multi-layered threat-modeling** strategy spanning three stages: (1) Capability assessment, where it maps plausible catastrophic-risk scenarios—actors, attack pathways, and harms—to determine whether model capabilities approach predefined Capability Thresholds; (2) Deployment safeguards, where it maps out the set of threats and vectors through which an adversary could catastrophically misuse the deployed system; (3) Security safeguards, where it seeks to establish the relationship between the identified threats, sensitive assets, attack vectors and, in doing so, sufficiently capture the resulting risks that must be addressed to protect model weights from theft attempts, using best practices such as the MITRE ATT&CK Framework. It does not mention the specific methodologies involved, lists of risk scenarios, and the complete risk models in the RSP.	The Framework identifies threat modeling as "a causal pathway for a severe harm in the capability area," which is one of the five criteria to meet to categorize a frontier risk to the Tracked Category. It is guided by both (1) the broader risk assessment process, and (2) more specific information that it gathers across OpenAI teams and external experts. The threat models are reviewed and approved by the internal, cross-functional group called Safety Advisory Group (SAG). It does not mention the specific methodologies involved, lists of risk scenarios, and the complete risk models in the Preparedness Framework.	The Framework describes risk modeling as "identifying and analyzing the main foreseeable paths through which a model could cause severe harm," and requires it for both risk assessment and mitigation assessment. The framework does not mention the specific methodologies involved, list of risk scenarios, and the complete risk models.	Meta's risk modeling exercises begin by testing whether the model has the (1) enabling capabilities and (2) could uniquely enable these scenarios to catastrophic outcomes. Inclusion for risk modeling follows a four-layered qualitative criteria, where risks have to be plausible, catastrophic, net new, and irreparable. The risk modeling process is informed by (1) internal assessment; (2) external engagements (governments, external experts, and the wider AI community). The qualitative risk scenarios are included in the risk threshold framework.	**The team adopts threat modeling specifically for Biological and Chemical Weapon risks.** Specifically, it breaks down the 5 critical steps where xAI models are restricted from providing detailed information or substantial assistance. These steps are defined qualitatively, in collaboration with external domain experts from organizations such as SecureBio, NIST, RAND, and EBRC. However, it does not construct specific risk scenarios combining some or all of these critical steps identified.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Indicator

Risk Analysis and Evaluation

Definition

This dimension assesses whether the company has established well-defined risk tolerances that precisely characterize acceptable risk levels for each identified risk. Moreover, this dimension examines if the company has successfully operationalized these tolerances into measurable criteria: Key Risk Indicators (KRIs) that signal when risks are approaching critical levels, and Key Control Indicators (KCIs) that demonstrate the effectiveness of mitigation measures. The assessment captures whether companies define these indicators in paired "if-then" relationships, where exceeding KRI thresholds triggers corresponding KCI requirements. This operationalization ensures that abstract risk tolerances translate into concrete, actionable metrics that guide day-to-day decisions and maintain risks within acceptable bounds.

Why This Matters

Without operationalizing risk tolerances into measurable metrics, companies cannot make consistent and evidence-based decisions about when to halt development or implement additional safeguards. Well-defined KRI-KCI pairs create accountability by establishing clear tripwires: when risk indicator X crosses threshold Y, control measure Z must be implemented. This systematic approach prevents ad-hoc decision-making during high-pressure situations and ensures that safety commitments translate into concrete actions rather than remaining aspirational statements.

Chinese Regulatory System Summary

Voluntary Technical Standard

The Risk Management Standard (Article 5.3.2) breaks down an organization's capability of risk analysis into three core components:

(1) classifying AI risks;

(2) analyzing the probability of AI risks, preferably through quantitative or semi-quantitative methods;

(3) analyzing the impact of AI risks, preferably through quantitative or semi-quantitative methods.

Moreover, Article 5.3.3 defines an organization's capability for risk evaluation as dependent on its ability to:

(1) Construct a probability-impact matrix;

(2) Prioritize risks accordingly, preferably combining quantitative and qualitative methods.

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
2.1 Setting a Risk Tolerance								
Measure 4.1 Signatories will establish clear and measurable thresholds for acceptable systemic risk for **each identified systemic risk**, informed by systemic-risk identification (Commitment 2) and analytical evidence from model data, evaluations, modeling, estimation, and post-market monitoring (Commitment 3). They will explain how these thresholds guide risk-acceptance decisions, justify why the approach ensures safety, and apply safety margins to account for uncertainty and potential mitigation failure.	The RSP has defined a **qualitative boundary of acceptable risks**, expressed as capability thresholds of the identified risks of CBRN weapons and autonomous AI R&D. These thresholds (CBRN-3, CBRN-4, AI R&D-4, AI R&D-5, and the autonomy checkpoint) marks the upper bound of risk that Anthropic considers acceptable to manage under existing deployment and security safeguards. Anthropic has not included how it has defined these thresholds and noted that they are "uncertain how to choose a specific threshold," but they "maintain a current list of specific CBRN capabilities of concern for which they would implement stronger mitigations," sharing only with selected organizations such as the AI Safety Institute and Frontier Model Forum.	The Framework establishes threshold levels of capability for when additional safeguards or no deployment apply. High and Critical capability thresholds refer to capabilities that increasing for severe harm in terms of existing and qualitatively new threat vectors respectively. For each risk in the Tracked Category, **capability thresholds qualitatively describe** things an AI system might be able to help someone do or might be able to do on its own that could meaningfully increase risk of severe harm, with corresponding threat models. OpenAI has not included how it has defined these thresholds.	The Framework establishes threshold levels of capabilities (CCLs) for when mitigation plans or suspension of deployment are required until risks are addressed. For each risk identified in the misuse category, **capability thresholds qualitatively describes how an AI system can "uplift" or autonomously carry out actions that will lead to risks of severe harm.** The CCLs are identified through "ongoing analysis" of the risk domains, which are expected by the team to "evolve over time," although the details of which are not included in the Framework. [Version 2.0 and Version 3]	The Framework establishes risk thresholds based on the extent to which a frontier AI model can uniquely enable execution of any of the threat scenarios. The framework introduces **a three-layered capability threshold** of **moderate, high, and critical**, which corresponds to (1) "release" - the model does not provide a significant uplift (2) "do not release" - the model can not yet uniquely enable a catastrophic threat scenario, but provides a significant uplift (3) "stop development" - the model can uniquely enable at least one complete catastrophic threat scenario	The RMF currently has sets **quantitative thresholds for Biological and Chemical risks**, which is to maintain an answer rate of less than 1 out of 20 on restricted queries; and for Loss of Control, which is to maintain a dishonesty rate of less than 1 out of 2 on MASK. It has cited plans to "add additional thresholds tied to other benchmarks." Performance against the Bio & Chem threshold is evaluated using an internal benchmark of benign and restricted biology- and chemistry-related questions developed in collaboration with SecureBio.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
2.2 Operationalizing Risk Tolerance								
	For each risk domain, two qualitative Key Risk Indicators (KRIs) are defined (CBRN-3, CBRN-4; AI R&D-4, AI R&D-5) to trigger escalation to ASL-3 or ASL-4 safeguards. **The indicators are primarily qualitative and not directly measurable**, with the exception of AI R&D-5, which specifies a quantitative benchmark based on effective scaling. No clear mapping is provided between these indicators and specific evaluation tests or quantitative thresholds, although Anthropic has noted that they prefer the flexibility of affirmative cases to board-approved evaluations. For each KRI, there are corresponding Key Control Indicators (KCIs) in the required safeguards that would apply upon escalation, including safeguards for deployment and security. These KCIs are defined qualitatively rather than quantitatively. The ASL-3 deployment safeguards "evaluate whether the measures Anthropic has implemented make us robust to persistent attempts to misuse the capability in question," but they do not include numerical thresholds or measurable performance criteria. The ASL-3 security safeguards "evaluate whether the measures Anthropic has implemented make us highly protected against most attackers' attempts at stealing model weights." The ASL-4 safeguards have not yet been defined.	For each risk domain, **two qualitative KRIs are defined**. The indicators are primarily qualitative, with the exception of AI R&D Critical, which specifies a more quantitative baseline. No clear mapping is provided between these indicators and specific evaluation tests or quantitative thresholds. For each KRI, there are corresponding KCIs in the required safeguards would apply upon escalation, including including security controls [High], safeguards against misuse [High], safeguards against misalignment [High], and development halts [Critical].	For **CBRN** and **Cyber risks**, the Framework defines qualitative thresholds for uplift capabilities. For Harmful Manipulation risk, an exploratory threshold is introduced. **ML R&D** risks now include two distinct thresholds: Acceleration Level 1, when models substantially accelerate AI progress beyond historical rates, and Automation Level 1, when models can fully automate the work of an AI research team. For **Misalignment** risks, the Framework retains two Instrumental Reasoning Levels as part of its exploratory approach For each KRI identified in the misuse risk categories, there exist corresponding KCIs as recommended security level (which is mapped to RAND Security Level) with the justifications. For the two instrument reasoning capabilities for misalignment risks, automated monitoring is required for level 1, while the team is still coming up with the approaches for Level 2.	**For each risk (cybersecurity and bio&chem weapons), 3 layers of qualitative catastrophic outcomes are identified**. For each outcome, 1-2 qualitative threat scenarios (Key Risk Indicators) are identified. Correspondingly, the threshold framework includes examples of model enabling capabilities for each threat scenarios. Meta deliberately withholds the detailed breakdown of how each threat scenario could be executed, citing concerns for balancing transparency vs. security. Meta does not include KCIs in accordance with KRIs.	**The quantitative threshold for malicious use risk and loss of control risk is not tied to any specific threat scenarios and is also not related to any specific safeguards accordingly**. While the RMF references safeguards at a high level, such as safety training, system prompts, and input & output filters, it does not specify how these measures are triggered, adjusted, or evaluated against the established thresholds.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Indicator

Risk Treatment

Definition

This dimension evaluates the extent to which the company has implemented comprehensive risk mitigation strategies across three critical areas: containment (controlling access to AI models), deployment (preventing misuse and accidental harms), and assurance processes (providing affirmative evidence of safety). Additionally, it assesses whether the company continuously monitors both key indicators throughout the AI system's lifecycle, from training through deployment.

Why This Matters

Effective risk treatment requires multiple layers of defense. Companies that maintain continuous monitoring of both risks and control effectiveness can detect when mitigations are failing before catastrophic outcomes occur.

Chinese Regulatory System Summary

Voluntary Technical Standard

The Risk Management Standard (Article 5.4) defines an organization's capability to handle risks based on two components:

(1) Selecting risk-response strategies;

(2) Developing and implementing risk-treatment plans, which preferably not only includes the ability to establish structured plans that specify responsibilities, timelines and priorities, but also ensure staff possess sufficient technical understanding and maintain effective, flexible, and timely execution.

The Risk Management Standard (Article 5.5) evaluates an organization's capability to monitor and review AI risks throughout the system's lifecycle. It consists of two main components:

(1) Risk Supervision which assesses whether whether the organization maintains continuous oversight of key risk areas—covering the supervision entity, scope of coverage, monitoring frequency, toolsets used, and response speed to emerging issues;

(2) Risk Inspection which is evaluated based on its coverage, timeliness, accuracy, practicality, and reliability.

Strategic and Policy Guidance Documents

The AI Safety Governance Framework 2.0 suggests strict control and full traceability of model applications to ensure that advanced AI systems cannot be exploited to develop or deploy large-scale lethal weapons. (Article 4.2.3(e))

Table begins on the next page

EU AI Code of Practice Safety and Security	**Anthropic**	**OpenAI**	**Google DeepMind**	**Meta**	**xAI**	**DeepSeek**	**Z.ai**	**Alibaba Cloud**
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
3.1 Mitigation Measures								
Measure 5.1 Signatories will implement safety mitigations that are appropriate along the entire model lifecycle, to ensure systemic risks stemming from the model are acceptable. (Commitment 4) **Commitment 6** Signatories will implement adequate cybersecurity protection for models and physical infrastructure along the entire lifecycle, to ensure systemic risks stemming from their models from unauthorised releases or access, and/or model theft are acceptable. **Measure 6.1** Signatories will define a goal that specifies the threat actors that their security mitigations are intended to protect against. **Measure 6.2** Signatories will implement appropriate security mitigations to meet the security goal, including the security mitigations pursuant to Appendix 4, such as general security mitigations, protection of unreleased model weights, hardening interface-access to unreleased model parameters, insider threats, and security assurance.	The latest model Sonnet 4.5 is deployed under ASL-3, according to the system card. In the Framework, ASL-3 Security Standards have clearly defined threat actors within scope and out of scope. It requires mitigation measures such as threat modeling, security framework, including parameters and access controls around sensitive assets, life cycle security, ongoing and effective monitoring, sufficient resourcing, and existing guidance, audits and documenting compliance when models are deployed in third-party environments. In addition, ASL-3 Deployment Standard requires threat modeling, defense in depth, red-teaming, rapid remediation, monitoring, trusted users, and documenting compliance when models are deployed in third-party environments. These measures are described as high-level outcomes and do not include actionable and measurable protocols.	The Preparedness Framework includes only illustrative examples of safeguards against malicious users, against a misaligned model, and security controls It also includes corresponding efficacy assessments for these safeguards. The latest model ChatGPT-5 is deployed under High Capability threshold for the Biological and Chemical Risks. The deployment includes multilayered mitigations—such as refusal and safe-completion training, real-time monitoring classifiers, account-level enforcement, and API safety identifiers.	The Framework does not specify the mitigation measures for the security controls at a level generally aligned with "security standards such as RAND SL 2, RAND SL 3, and RAND SL 4." It explained such decisions are due to the fact that they "expect the concrete [security] measures implemented to reach each level of security to evolve substantially." Deployment mitigations involve processes that are "designed to ensure that residual risk remains at acceptable levels," which involves (1) the development and assessment of mitigations; (2) pre-deployment review of safety case; (3) post-deployment where safety cases and mitigations may be updated if deemed necessary by post-market monitoring. The latest model Gemini 2.5 Pro did not reach the CBRN Uplift Level 1 CCL, Cyber Autonomy Level 1 and Uplift Level 1, Machine Learning R&D Uplift Level 1 and Autonomy Level 1. However, alert thresholds for the model's alert threshold for Cyber Uplift Level 1 prompted proactive measures—specifically, increased evaluation cadence and accelerated mitigation deployment.	The Framework states that the full mitigation strategy will be informed by the risk assessment, the frontier AI's particular capabilities, and the release plans. It does not prescribe a fixed set of mitigations, but list a few examples include certain examples including fine-tuning, misuse filtering, response protocols, sanctions screening and geo-gating, staged release to prepare the external ecosystem. Meta has not updated Llama 4 Maverick's system cards to reflect these changes.	The RMF references mitigations measures on a high level, including: (1) safety training, system prompts, and input & output filters for malicious use risks (2) safety training for controllability, and system prompt for loss of control risks. These mitigations do not correlate with the aforementioned threshold. The latest model Grok-4 have implemented safeguards in particular for the Bio & Chem risk, including (1) narrow, topically-focused filters for bioweapon abuses and chemical weapons-related abuses; (2) existing system prompts against radiological and nuclear weapons development.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
3.2 Continuous Monitoring and Comparing Results with Predetermined Thresholds								
	Anthropic's capability assessment for the most pressing risks has three stages: (1) preliminary testing, (2) comprehensive evaluation, and (3) a capability decision. - Models showing (1) a 4× increase in Effective Compute or (2) six months of fine-tuning trigger full testing. - Comprehensive evaluation covers threat modeling, empirical testing, elicitation under attacker scenarios, and forecasting. - Results are reviewed by the Responsible Scaling Officer (RSO) and CEO to decide if escalation is needed. Accordingly, it also assesses the safeguards of the Deployment and Security Standard. After evaluating their implementations, the CEO and the RSO (1) make the ultimate determination as to whether we have satisfied the Required Safeguards and (2) decide any deployment-related issues after soliciting internal and external expert feedback on the evaluation. Safeguards will be revisited and re-approved at least annually. There also exists follow-up capability assessment that rechecks to ensure model capabilities remain below higher thresholds and updates the policy if new risks emerge.	Before deployment, every model covered by the Framework undergoes a structured suite of Scalable Evaluations: automated tests that measure capability proxies tied to risk thresholds. The results of which will be compiled into a Capabilities Report that is submitted to the SAG. The report will be reviewed by the SAG to decide on the next steps, which can include (1) Capability threshold is crossed, recommending to implement and assess corresponding safeguards; (2) Capability threshold has not been met, (3) Recommend deep Dive evaluations, such as expert red-teaming or third-party assessments, to validate those results. Accordingly, it also assesses the safeguards through a Safeguards Report, which compiles all identified pathways by which severe harm could occur, the corresponding mitigations, their measured efficacy, the residual risk after controls are applied, and notable limitations. The SAG reviews this report to determine whether the safeguards in place sufficiently minimize the risks associated with the model's capability level and deployment context, drawing on internal and external expert input as needed.	Google conducts regular early-warning evaluations to monitor whether models are nearing critical capability levels. These evaluations use predefined alert thresholds and are adjusted in frequency or sensitivity as model progress accelerates. When needed, they are supplemented by additional assessments to ensure an accurate understanding of capability proximity and to maintain a sufficient safety buffer before deployment.	Meta lays out the monitoring and continuous evaluation process in the following procedure: (1) Ensuring robust evaluation environment (2) Conducting evaluations for performance and safety, against our expectations for the reference class as well as the enabling capabilities we have identified in our threat scenarios. - The indicators trigger for further evaluations as capabilities develop. (3) Evaluations are repeated as a frontier model is close to or completes training.	xAI continuously measures model's safety properties through public benchmarks and monitors live use through public deployment (e.g. Grok on X) It also regularly evaluates the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that we could potentially utilize, to determine and apply effective benchmarks available at the time of evaluation.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Indicator

Risk Governance

Definition

This dimension examines whether the company has built robust organizational infrastructure to support effective risk management decision-making. The assessment captures the extent to which companies have established clear risk ownership and accountability, independent oversight mechanisms, and cultures that prioritize safety alongside innovation. Moreover, this dimension evaluates the company's commitment to transparency, specifically their public disclosure of risk management approaches, governance structures, and safety incidents. The evaluation considers how well the company's governance framework ensures that risk considerations are incorporated into strategic decisions and that multiple layers of review prevent any single point of failure in risk management.

Why This Matters

Strong governance structures ensure that risk management isn't just a technical exercise but is embedded in organizational decision-making at all levels. Independent oversight prevents conflicts of interest when safety considerations clash with commercial pressures, while clear accountability ensures someone is always responsible for catching problems. Companies that publicly disclose their governance structures and safety incidents demonstrate confidence in their approach and enable external stakeholders to verify that appropriate safeguards exist.

Chinese Regulatory System Summary

Local Binding Instruments

Shanghai Regulation (2022) requires that the high-risk AI products and services be subject to list-based management and undergo compliance review in accordance with the principles of necessity, legitimacy, and controllability. (Article 65)

Shenzhen Regulation (2022) requires the high-risk AI applications to adopt a regulatory model of ex-ante assessment and risk warning. (Article 66) These two regulations do not apply to Z.ai (Beijing), DeepSeek (Zhejiang), or Alibaba (Zhejiang).

Voluntary Technical Standard

The Risk Management Standard (Article 5.1) evaluates an organization's ability to plan and organize AI risk management activities, including:

(1) Leadership and Governance (Article 5.1.1)— assessing whether senior leadership establishes clear organizational policies and objectives for AI risk management, allocates sufficient resources, and assigns defined responsibilities.

(2) Policy Development (Article 5.1.2) — examining whether the organization defines the scope of AI risk management, sets parameters and evaluation criteria, and establishes consistent strategies and resource reserves for managing risks.

Strategic and Policy Guidance Documents

Ethical Norms for New Generation Artificial Intelligence (2021) establishes that all types of AI activities shall comply with the basic ethical norms listed in this document, which include Assurance of Controllability and Trustworthiness. This means ensuring that humans have fully autonomous decision-making rights and that they have the right to accept or reject AI-provided services, the right to withdraw from AI interactions at any time, and the right to terminate AI system operations at any time. Ensure that AI is always under human control. (Article 3)

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
4.1 Decision Making								
Measure 4.2 Signatories will base go/no-go decisions for model development, release, and use on whether systemic risks are deemed acceptable (Measure 4.1). **Measure 8.1** Signatories will clearly define, assign and document systemic-risk responsibilities across all organizational levels, including systemic risk oversight, ownership, support and monitoring, as well as assurance. **Measure 8.2** Those who have been assigned responsibilities (Measure 8.1) should be allocated appropriate human, financial and computational resources as well as access to information.	**Go/no-go decisions made by the CEO and RSO** based on whether risks and safeguards remain acceptable under ASL thresholds. These decisions then escalate to the Board of Directors and the Long-Term Benefit Trust before moving forward.	**The Safety Advisory Group (SAG) makes expert recommendations on whether safeguards are sufficient for deployment**; however, OpenAI Leadership can approve or reject these recommendations, and the Board's Safety and Security Committee provides oversight of these decisions.	Response plan to when alert thresholds are reached will be reviewed and approved by **appropriate corporate governance bodies**, such as (1) Google DeepMind AGI Safety Council, (2) Google DeepMind Responsibility and (3) Safety Council, and/or Google Trust & Compliance Council. [Version 2.0]	After the continuous evaluation process, the team will conduct residual risk assessments, which is informed by evaluations and mitigations. The results are reviewed by research and product teams and a multidisciplinary review group (as needed). A leadership team will then decide whether to approve, require further testing, or halt release, guided by the risk thresholds.	Deployment is gated by benchmark-linked thresholds and a tiered-access strategy; functionality can be restricted to only trusted parties. Where warranted, xAI may revoke accounts, temporarily shut down systems, or notify authorities to prevent materially unjustified risk increases. The RMF does not explicitly define how deployment decisions are reached, arguing that "the expected benefits of model deployment may outweigh the risks identified by a particular benchmark," suggesting that risk assessment and capability evaluation results may not automatically trigger decision to pause development and stop deployment.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.2 Advisory and Challenge								
Measure 8.1 Signatories will designate at least one member from the management body to support and monitor systemic-risk management, including conducting risk assessments and mitigations	RSO is designated to be responsible for reducing catastrophic risk, primarily by ensuring that the policy is designed and implemented effectively. Its specific duties are also clearly defined, covering the full life stages of policy development to policy enforcement.	The SAG is the internal cross-functional advisory body that reviews threat models, Capability Reports, Safeguards Reports and makes recommendations to OpenAI Leadership regarding the level and type of safeguards required for deploying frontier capabilities safely and securely.	The DeepMind AGI Safety Council will periodically review the implementation of the Framework. [Version 2.0]	It is unclear which leadership team will be responsible for supporting and monitoring the systemic risk management.	No internal body has been appointed or identified to support and monitor the systemic risk management. But the RMF integrates the approach of designating risk owners, who are responsible also for proactively mitigating identified risks.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
4.3 Audit								
Measure 8.1 Signatories will designate an assurance role (e.g., Chief Audit Executive or Head of Internal Audit) that is tasked with providing assurance on the adequacy of systemic-risk processes to the board or its supervisory function. This individual is supported by internal audit and, where appropriate, external auditors.	ASL-3 Security requires the mechanism to (1) audit and assess the design and implementation of the security program and (2) share these findings with management on an appropriate cadence. The following methods have been recommended: independent validation of threat modeling and risk assessment results; a sampling-based audit of the operating effectiveness of the defined controls; periodic, broadly scoped, and independent testing with expert red-teamers who are industry-renowned and have been recognized in competitive challenges.	The framework requires auditing and transparency mechanisms as part of the security controls for High capability models. These measures include independent security audits to security controls and practices are validated regularly by third-party auditors to ensure compliance with relevant standards and robustness against identified threats.	Auditing was mentioned as an example of the suite of safeguards targeting the capability, although it is not a formal part of the deployment mitigations.	There is no mention of internal or external audit functions in the Framework.	There is no mention of internal or external audit functions in the Framework.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.4 Oversight								
Measure 8.1 Signatories will assign a specific committee of the management body in its supervisory function or one or more multiple suitable independent bodies to oversee its systemic risk management processes and measures.	**Oversight is provided by the Board of Directors**, including the Long-Term Benefit Trust, which review risk determinations, safeguard implementation, and deployment decisions under the RSP.	**Oversight is provided by the Board's Safety & Security** Committee, which receives information on process and decisions and "may reverse a decision or mandate a revised course of action" if necessary.	Appropriate corporate governance bodies such as the Google DeepMind AGI Safety Council, Google DeepMind Responsibility and Safety Council, and/or Google Trust & Compliance Council will review and approve response plans, while Google DeepMind AGI Safety Council will periodically review the implementation. [Version 2.0]	A leadership team will then decide whether to approve, require further testing, or halt release, guided by the risk thresholds, although it is unclear who will make up the leadership team.	No oversight body has been identified in the RMF.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
4.5 Culture								
Measure 8.3 Signatories will promote a healthy risk culture and take appropriate measures to ensure that actors who have been assigned responsibilities for managing the systemic risks stemming from their models (**Measure 8.1**) take a reasoned and balanced approach to systemic risk. Examples include leadership priority, clear communication and challenge of decisions concerning systemic risks, active internal reporting channels, no retaliation, incentives and structural independence for objective risk assessment and less excessive risk-taking, and easy public access and regular reminder of whistleblower policy.	Anthropic protects employees' ability to raise safety and compliance concerns without retaliation by maintaining anonymous reporting channels for noncompliance to the RSO and the Board of Directors and prohibiting non-disparagement clauses that could discourage speaking up about safety issues. Anthropic has multiple teams working on AI safety research including alignment science, interpretability, frontier red team, safeguards team and more.	OpenAI's employees can access summaries of Safety Advisory Group (SAG) testing results and recommendations, within confidentiality limits. All potential policy violations or implementation issues can be reported under the Raising Concerns Policy, and each report is tracked, investigated, and addressed with proportional corrective actions. (The whistleblower policy will be discussed more in detail in "Governance and Accountability" Section).	No internal reporting or anti-retaliation mechanisms are referenced in the Framework.	No internal reporting or anti-retaliation mechanisms are referenced in the Framework.	Employees can raise concerns to relevant government agencies regarding imminent threats to public safety based on whistleblower policy.		Z.ai's safety team is made up of Zhipu Evaluation Team, Zhipu Safety Team, Zhipu Posttraining Team. The teams do not have team websites and prefer not to disclose mission and scope. There are 20-30 technical FTEs for safety teams.	
4.6 Transparency								
Commitment 7 Safety and Security Model Reports Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office. - Content Requirements (Measure 7.1-Measure 7.5), such as model description and behavior, reasons for proceeding with development, documentation of risk identification, analysis, and mitigation, external reports, and material changes to the systemic risk landscape	**Anthropic promises to share publicly key information** related to the evaluation and deployment, including (1) Capability and Safeguards Reports for deployed models, (2) plans for comprehensive capability assessments and deployment and security safeguards. **It will also ask for external input from experts** for developing and conducting the capability and safeguards assessments and third-party review of procedural commitments on an approximately annual basis.	**OpenAI promises to share with the public** summaries of capability evaluations, testing scope, reasoning behind deployment decisions, and implemented safeguards (for models at or beyond the High threshold), with redactions where needed for security or proprietary reasons.	The Frontier Safety Framework will be updated at least once a year, including the CCLs and the testing and mitigation approaches.		xAI intends to publish publicly and for third-party reviews with potentially redacted information for concerns of public safety, national security, and protection of intellectual property: (1) Updates to the RMF (2) Adherence with the RMF (3) Benchmark results (4) Internal AI Usage (5) Employee survey for important future developments of AI		Z.ai has a written formal policy to conduct regulator-only notification, where the policy mandates prompt disclosure to a competent regulatory, or supervisory authority when safety testing determines a model exceeds its "unacceptable-risk" threshold.	

Table continues on next page

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
- Update Duties when signatories have reasonable grounds to believe if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable - Notifications **Measure 10.1** Signatories must maintain comprehensive internal documentation on model architecture, system integration, evaluations, and safety mitigations. They must also record processes, key risk-related decisions, and justifications for their chosen safety practices. Documentation must be kept for at least 10 years and be made available to the AI Office upon request. **Measure 1.3** Signatories will update the Framework as appropriate, including without undue delay after a Framework assessment to ensure the information for the safety framework is kept up-to-date and the Framework is at least state-of-the-art. For any update of the Framework, Signatories will include a changelog, describing how and why the Framework has been updated, along with a version number and the date of change. Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office.	**The company will also notify U.S. government authorities** if stronger protections than ASL-2 are needed. In the system card for Sonnet 4.5, Anthropic has noted that the model does not require comprehensive capability assessment since it does not meet the "notably more capable" threshold. Comprehensive automated testing, comparative capability assessment to earlier models, and conservative threshold application evaluations confidently rule out ASL-4 capabilities across all domains. The decision was overseen by the RSO and followed the company's established protocols for precautionary ASL determinations	When warranted, OpenAI will engage independent third parties to evaluate model capabilities and stress-test safeguards, particularly for high-risk deployments. The SAG may also seek independent expert opinions to inform its safety determinations before deployment. In the system card for GPT-5, OpenAI recorded both scalable and deep-dive evaluations for the model across the three Tracked Categories, including **both internal and external assessments** compiled into a Capabilities Report for the SAG. The SAG reviewed the evidence and concluded that GPT-5-Thinking reached the High threshold, requiring "safeguards sufficiently minimize associated risks" before deployment. The Preparedness Team compiled mitigations into a Safeguards Report, **validated through extensive third-party red-teaming**. The SAG, supported by OpenAI leadership and external experts, provided oversight across the evaluation and mitigation phases. **There is no written requirement to notify any external body if safety testing determines a model exceeds OpenAI's "unacceptable-risk" threshold.**	**Google DeepMind is dedicated to sharing relevant information with appropriate government authorities when a model has reached a CCL according to their assessments**. These disclosures occur under strict confidentiality and security safeguards. Such information may include model information, evaluation results, and mitigation plans. Google DeepMind also considers disclosing information to other external organizations to promote shared learning and coordinated risk mitigation, although unclear under what circumstances.	In the Framework, Meta states their continuous dedication to openly releasing models to the ecosystem, sharing relevant information about responsible development and evaluation through model cards and research papers and believes that this will allow their team to work with outside experts and allow external independent assessment of their models. However, according to a letter released by Mark Zuckerberg on July 30, 2025, the CEO of Meta noted that the company will be "careful about what we choose to open source."			

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Safety Frameworks

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Comprehensive framework with clear systemic-risk identification, modeling, thresholds, mitigations, and governance; strong accountability and documentation.
- **B** Robust framework that covers key systemic-risk areas with defined thresholds and oversight; minor gaps in scope or clarity.
- **C** Basic framework; outlines risk areas and mitigations but lacks clear thresholds or governance detail.
- **D** Weak framework; vague risk identification and mitigations; governance and accountability poorly defined.
- **F** No credible framework; systemic risks, mitigations, and governance absent.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Existential Safety

This domain examines companies' preparedness for managing extreme risks from future AI systems that could match or exceed human capabilities, including stated strategies and research for alignment and control.

Table of Contents

Chinese Regulatory System Summary

Speech by high-level government leadership and recent governance frameworks have indicated broad direction for future AI regulation, focusing on preventing “loss of control” risks of frontier AI systems, and ensuring that AI systems remain under human control.

Strategic and Policy Guidance Documents

AI Safety Governance Framework 2.0 emphasizes the principles of “safety, reliability, and controllability” for AI development, to “strictly prevent loss of control risks that could threaten the survival and development of humanity, and to ensure that AI is always under human control.” (Article 1.5)

Li Qiang (Premier) "No matter how technology transforms, it must remain a tool to be harnessed and controlled by humans. AI should become an international public good that benefits humanity." (July 2025)

Xi Jinping "urged efforts to consistently strengthen basic research and focus on overcoming challenges regarding core technologies such as high-end chips and foundational software, thereby building an independent, controllable, and collaboratively-functioning foundational software and hardware system for AI." (April 2025)

Indicator

Existential Safety Strategy

Definition

The assessed companies aim to develop AGI/superintelligence, and many expect to achieve this goal in the next 2–5 years. This indicator evaluates whether companies have published comprehensive, concrete strategies for managing catastrophic risks from these transformative AI systems. We assess the depth, specificity, and credibility of publicly available plans. We examine official company documents, research papers, and blog posts that articulate safety strategies. We report the most relevant documents, briefly summarize their content, and provide links for detailed reading. Safety frameworks are mentioned for completeness and are fully evaluated in the relevant domain. We note whether documents are declared strategies by leadership or proposals by researchers from a safety team. We strive to keep document summaries proportional to document length and relevance for the safety strategy. Safety frameworks are only noted briefly and evaluated in another domain. Documents that primarily provide recommendations to other actors (e.g., governments) are outside the scope.

Key components:

Technical Alignment and Control Plan:

- Given the short timelines to AGI and the magnitude of the risk, companies should ideally have credible, detailed agendas that are highly likely to solve the core alignment and control problems for AGI/Superintelligence very soon.
- Companies should be able to demonstrate that they would be able to detect misaligned systems and reliably prevent them from escaping human control, and have formulated clear protocols for how they will handle serious warning signs of misalignment.

AGI Planning:

- Companies should have detailed plans for managing the transition when AI matches or exceeds human capabilities in critical domains and enables large scale dual-use risks. They should specify clear criteria for when they would halt development/deployment.
- Companies should develop concrete, detailed roadmaps to achieve sufficient cyber-defense capabilities to protect against attacks from terrorist organizations or resourced state actors before critically dangerous systems are developed.

Post-AGI Governance:

- Companies should provide clear descriptions of how they would govern AGI/Superintelligence or how they will enable societal control. The company also should have developed reliable protocols that would prevent insiders from using superintelligent systems to seize political power.
- Companies should specify how extreme power concentration will be prevented and benefits distributed if AI replaces humans in the workplace and causes unprecedented mass unemployment.

Overall, this indicator evaluates whether companies have detailed, actionable strategies that match the extraordinary risks they acknowledge when building systems intended to exceed human intelligence.

Why This Matters

Industry leaders and the recent International Scientific Report on the Safety of Advanced AI have identified potentially catastrophic risks from advanced AI systems. Several assessed companies predict AGI development within 2-5 years, creating urgency for reliability, safety preparedness. This indicator summarizes core documents that are relevant to a company's posture toward these risks. Given the irreversible nature of potential failures and their global impact, the sophistication of a company's strategy should scale with its stated ambitions and timelines. A well-defined existential safety strategy, backed by clear governance, resources, step-by-step implementation, and transparency, signals readiness to act responsibly in managing civilization-scale risks.

Table begins on the next page

Company Strategy		Quantitative Safety Plan (quantitative bounds on control/ alignment failure risk)
Anthropic	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Update** No notable AGI strategy updates since May 2025. **Recap from Summer 2025** **Foundational philosophy & Long-term scenarios** In "Core Views on AI Safety" (2023), Anthropic has laid out three possible futures (optimistic, intermediate, and pessimistic) depending on how tractable alignment proves to be. It also identified 6 long-term research pillars: Mechanistic Interpretability, Scalable Oversight, Process-Oriented Learning, Understanding Generalization, Testing Dangerous Failure Modes, and Societal Impact Evaluation. **Foundational governance structure** Anthropic has continuously updated its Responsible Scaling Policy, including the most recent updates in May 2025, to publicize its commitment to pausing model training or deployment if systems reach predefined Capability Thresholds without safety and adequate safeguards. The policy institutionalizes internal oversight through a Responsible Scaling Officer and the Board, mandatory risk assessments, and incident readiness exercises. **Research Agenda** The team has continued to emphasize research effort to manage rapidly advancing model capabilities. In "The Urgency of Interpretability" (2025), CEO Dario Amodei positions interpretability research as a race against accelerating intelligence, aiming by 2027 for tools that can "reliably detect most model problems." Complementing this, Sam Bowman's "Putting up Bumpers" (2025) advances an engineering-based alignment approach built on continuous testing and overlapping safety mechanisms.	No public-facing quantitative safety plan found
OpenAI	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Update** In "Security on the Path to AGI," [1], OpenAI has shared their security initiatives on advancing to AGI, including an expanded Cybersecurity Grant Program and Bug Bounty Program, partnerships for continuous adversarial red teaming, deployment of AI-powered cyber defense systems, stronger safeguards for advanced AI agents such as Operator and Stargate, and adoption of zero-trust, hardware-backed infrastructure to scale security alongside advancing model capabilities. **Research Agenda** The company believes in avoiding optimization that encourages obfuscation: Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior. In the company survey, the company stated that "We've published research and joined a broader working paper urging against optimizing on chains of thought: As we noted in the GPT-5 system card, "our commitment to keep our reasoning models' CoTs as monitorable as possible (i.e., as faithful and legible as possible) allows us to conduct studies into our reasoning models' behavior by monitoring their CoTs." **Recap from Summer 2025** **Foundational philosophy and strategy** OpenAI stated in its strategy "How we think about safety and alignment," that it has shifted from viewing AGI as a single transformative moment to seeing it as continuous progress. It further listed its core principles that currently guide the company's thinking and actions, which include Embracing uncertainty, Defense in Depth, Methods that Scale, Human Control, and Community Effort. For every principle, the blog lays out how it will shape their focus and approach to new challenges and relates to already implemented interventions. This thinking iterates on the 2023 blog post "Planning for AGI and beyond," emphasizing goals including ensuring AGI benefits are "widely and fairly shared" and advocates for deploying progressively more powerful systems to learn iteratively. **Foundational governance structure** Preparedness Framework, which is updated in April 2025, describes OpenAI's commitment to pausing development or deployment if required mitigations cannot adequately address the identified risks based on regular dangerous capability evaluations and the predefined capability threshold triggers.	No public-facing quantitative safety plan found

Table continues on next page

Company Strategy		Quantitative Safety Plan (quantitative bounds on control/ alignment failure risk)
Google DeepMind	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Update** Google DeepMind has updated its Frontier Safety Framework in September 2025. Compared to v2.0, the updated version introduced new risk domain (harmful manipulation) in the misuse risk section, broadening the section on misalignment risks from deception, increased transparency on external disclosure, and expand mitigation coverage to large-scale internal deployment. **Recap from Summer 2025** **Research agenda and efforts** An Approach to Technical AGI Safety and Security (April 2025) A detailed technical report by DeepMind's safety team explains their research agenda for preventing severe, civilisation-scale harm from AGI—defined as systems roughly at the 99th-percentile of skilled adults. The paper identifies four areas of risk: misuse, misalignment, mistakes, and structural risks and chooses to focus on technical approaches to misuse and misalignment. The strategy for misuse is to proactively identify dangerous capabilities and implement robust security, access restrictions, monitoring, and model safety mitigations to prevent threat actors from accessing these dangerous capabilities. The strategy for misalignment is "two lines of defense," including model-level mitigations + system-level security measures. The safety-case methodology serves as the integrative layer connecting these safeguards, as it proposes making deployment decisions through structured, evidence-based arguments: inability cases (model lacks capability) and control cases (misaligned behaviour will be caught). **Foundational governance structure** Frontier Safety Framework (v 2.0) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations. These commitments include pausing development or deployment if the required mitigations cannot adequately manage the identified risks.	No public-facing quantitative safety plan found
Meta	**No existential safety strategy found, but evidence below that the company has started to engage with the topic.** **Update** Meta's Shift on Open-Source AI In July 2025, Mark Zuckerberg wrote in a blog post "Personal Superintelligence," that Meta "will need to be rigorous about mitigating these risks and careful about what [it] choose to open source. Still, [Meta] believe that building a free society requires that [it] aim to empower people as much as possible." **Recap from Summer 2025** **Foundational philosophy** Open Source AI Is the Path Forward (2024) In this blog post, Zuckerberg presents a case for open source AI as their primary approach to AI safety and development (not specifically focused on catastrophic risks). The document makes the case that open source models are inherently safer than closed alternatives due to transparency, distributed scrutiny, and prevention of power concentration. **Foundational governance structure** Frontier AI Framework v.1.1 (2025) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations. These commitments include pausing development or deployment if the required mitigations cannot adequately manage the identified risks.	No public-facing quantitative safety plan found

Table continues on next page

Company	Strategy	Quantitative Safety Plan (quantitative bounds on control/ alignment failure risk)
xAI	**No existential safety strategy found, but evidence below that the company has started to engage with the topic.** **Update** xAI Risk Management Framework (August 2025) The formalized RMF outlines xAI's approach to policies for handling significant risks associated with the development, deployment, and release of AI models such as Grok. It identifies quantitative thresholds and metrics for a few critical risks, and lays out procedures that could be used to manage and improve the safety of AI systems. **Recap from Summer 2025** **Foundational governance structure** xAI Risk Management Framework (Draft) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations.	No public-facing quantitative safety plan found
DeepSeek	**No public-facing existential risk policy found**	No public-facing quantitative safety plan found
Z.ai	**The company has indicated in the company survey that it doesn't yet have an AGI explicit existential risk strategy, but is actively developing one.**	No public-facing quantitative safety plan found
Alibaba Cloud	**No public-facing existential risk policy found**	No public-facing quantitative safety plan found

Footnotes

[1] OpenAI has included the link to this blog post in the company survey to provide "additional information about our security work that it believes may be useful context for evaluators considering its overall posture and approach" as part of their strategies towards safe and controllable AGI.

Indicator

Internal Monitoring and Control Interventions

Definition

This indicator evaluates whether companies have implemented or prepared monitoring and control systems to detect and prevent risks from misalignment during internal deployment. Companies are assessed on whether they have concrete implementation plans tied to specific capability thresholds, published methodologies for control evaluations, and protocols for investigating potential scheming or deceptive alignment. General statements about monitoring without specific technical details, thresholds, or implementation timelines are insufficient. Research about monitoring without statements on implementation plans or status is out of scope.

Why This Matters

As AI systems become more capable, they may develop the ability to engage in deceptive behavior or "scheming"—appearing aligned while pursuing misaligned goals that could include attempts to gain unauthorized access to resources, sabotage safety research, subvert oversight mechanisms, or manipulate staff. Internal deployment poses unique risks, as this is usually the first time a highly capable AI system has longer time-horizon interactions with the external world. Robust monitoring and control measures after deployment serve as a critical line of defense, enabling companies to detect and prevent harmful actions even if alignment techniques fail to prevent scheming entirely. Concrete and technically sophisticated control protocols indicate that companies are taking tail risks seriously and work on mitigations.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
No public control plan found. The Responsible Scaling Policy says that when AI can fully automate the work of junior AI researchers, they "will develop an affirmative case that (1) identifies the most immediate and relevant risks from models pursuing misaligned goals and (2) explains how we have mitigated these risks to acceptable levels. The affirmative case will describe [..] mitigations (such as monitoring [..] [Responsible Scaling Policy, 2025].	**No public control plan found.** Preparedness Framework states that a "High" risk classification for "Cybersecurity" domain triggers requirement for ".. safeguards meeting High standard against misalignment (Appendix C.2) for large-scale internal deployment". Appendix C.2 contains illustrative safeguards against misaligned models. Sub-section on "Reliable and Robust System Oversight" lists 4 potential measures: - Monitor AI supervises agent actions to detect agent in the case of trying to take harmful actions, scheming, or similar - System forces human approval of agent actions (either by default, or by escalation of the monitor) - Model's reasoning is provided faithfully and interpretably to humans to review - Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm [OpenAI, 2025]	**Early sketch of a plan exists, without technical details.** From Frontier Safety Framework (v3.0): "An initial mitigation approach focuses on detecting when models might develop a baseline instrumental reasoning ability at which they have the potential to undermine human control, assuming no additional mitigations were applied. When models reach this capability level, we believe applying an automated monitor to the model's explicit reasoning (e.g. chain-of-thought output) is an effective mitigation. Once a model is capable of effective instrumental reasoning in ways that cannot be monitored, additional mitigations may be warranted—the development of which is an area of active research." The previous iteration of the Framework (v2.0) states that developers should implement 'control evaluations' [Greenblatt, 2023] and construct a safety case showing: "that the likelihood of deceptive alignment risk would be low enough for safe deployment, even if the model were trying to meaningfully undermine human control."	**No public control plan found.**	**No public control plan found.**	**No public control plan found.**	The company has indicated in its survey response that it maintains control interventions around emergency response and has demonstrated internal monitoring readiness, **although no formal or publicly available plan has been disclosed.** **(1) Control interventions** The company maintains multiple mechanisms designed to enable rapid containment and mitigation of safety incidents, including i) technical capability to rapidly roll back a deployed model to a previous version globally (within 12h), ii) technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g., web-browsing) globally **(2) Monitoring readiness** It has i) conducted at least one full live emergency response drill/ simulation in the past 12 months, and has ii) created a formal and documented emergency response plan for AI safety incidents that delineates trigger threshold, named incident commander, and 24*7 duty roster.	**No public control plan found.**

Indicator

Technical AI Safety Research

Definition

This indicator tracks AI company's research publications on technical AI safety research that are relevant to extreme risks. More specifically, the indicator is a collection of work that is plausibly helpful for averting large-scale risks from misalignment or misuse. This includes mechanistic interpretability, scalable oversight, unlearning, model organisms of misalignment, model evaluations on dangerous capabilities or alignment, and others. The collection also includes substantial outputs besides papers—weights, tools, code, transcripts, data—but these are almost always published as part of a paper. Excluded are capability-focused research, papers on hallucinations, model cards.

The full collection was created by Zach Stein-Perlman as part of his efforts at AI Lab Watch to evaluate company's practices of boosting safety research. His dataset covers publications up to July 2025. we have extended it to include works released through November 8, 2025, and added entries for DeepSeek, Zai, xAI, and Alibaba, based on additional research by the FLI team.

Why it matters

The industry is rapidly advancing toward increasingly capable AI systems, yet core challenges—such as alignment, control, interpretability, and robustness—remain unresolved, with system complexity growing year by year. Safety research conducted by companies reflects a meaningful investment in understanding and mitigating these risks. When companies publicly share their safety findings, they enable external scrutiny, strengthen the broader field's understanding of critical issues, and signal a commitment to safety that goes beyond proprietary interests.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Total	34	19	30	6	0	0	0	2
2025	9	-	4	1	3	0	-	1
2024	11	-	11	5	7	0	-	1
2023	12	-	13	-	2	0	-	-

Indicator

Supporting External Safety Research

Definition

This indicator assesses the extent to which companies invest in and support external AI safety research through a range of mechanisms. Evidence may include: (1) Mentorship programs—participation in formal initiatives such as the Machine Learning Alignment Theory Scholars (MATS) program, the number of mentors provided, and the existence of company-specific fellowships; (2) Research grants and funding—provision of financial support or subsidized API access to safety researchers, including grants and targeted funding programs; and (3) Deep model access for safety researchers—offering privileged access that goes beyond public APIs, such as employee-level permissions, early access to unreleased models, safety-mitigation-free versions for testing, fine-tuning rights on frontier AI systems, and allocated compute resources.

Why This Matters

External safety researchers often lack the access or funding to do the most valuable work they can. Companies committed to ecosystem-wide safety progress should empower the research community by providing deeper access to frontier AI systems, mentoring the next generation of research talent, and supporting funding-constrained external researchers. Deep model access enables critical research into the true model capabilities, alignment properties, and internal workings. Company-provided compute resources and API credits can help academics and independent researchers with limited financial resources to experiment on frontier models.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
AI Safety researcher Ryan Greenblatt from Redwood Research was given employee-level access in 2024, leading to the published research titled "Alignment Faking in Large Language Models."	Non-frontier model gpt-oss-120b and gpt-oss-20b model weights publicly available	Non-frontier model Gemma 3 model weights publicly available	Frontier model weights are publicly available	Non-frontier model Grok-1 model weights are publicly available	Frontier model weights are publicly available	Frontier model weights are publicly available	Non-frontier model Qwen3 model weights are publicly available

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Existential Safety

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Comprehensive, evidence-based strategy with quantitative safeguards and research plans for alignment and loss-of-control prevention.
- **B** Strong strategy; clear alignment objectives and technical pathways likely to prevent catastrophic risks.
- **C** Basic strategy; general preparedness and research focus with limited technical or measurable safeguards.
- **D** Weak strategy; vague or incomplete plans for alignment and control; minimal evidence of technical rigor.
- **F** No credible strategy; lacks safeguards or increases catastrophic-risk exposure.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Governance & Accountability

This domain audits whether each company's governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems.

Table of Contents

Chinese Regulatory System Summary

China does not have a regulatory framework for protecting whistleblowers, especially in the area of AI safety.

Indicator

Company Structure & Mandate

Definition

This indicator evaluates whether a company's fundamental legal structure, ownership model, and fiduciary obligations enable safety prioritization over short-term financial pressures in high-stakes situations. We report any embedded durable commitments to safety, social welfare, and benefit sharing and focus on any legally binding mechanisms (e.g., PBC status, capped equity, empowered governance bodies) that constrain management or shareholder incentives.

Why This Matters

Structural governance commitments can influence how companies respond when safety considerations conflict with profit incentives. During competitive pressures or deployment races, traditional for-profit structures may legally compel management to prioritize shareholder returns even when activities may pose significant societal risks. Structural governance innovations that formally embed safety into fiduciary duties—such as Public Benefit Corporation status or capped-profit models—create legally binding constraints that can override short-term financial pressures.

Anthropic	**Same as AI Safety Index Summer 2025** Uncommon governance structure. Finetuned for the ability to handle extreme events with humanity's interests in mind. Delaware Public Benefit Corporation (PBC) with a public benefit purpose. Anthropic's Purpose: "responsible development and maintenance of advanced AI for the long-term benefit of humanity." The Long-Term Benefit Trust (LTBT) is an independent body of five financially disinterested members, with the same purpose as PBC. It has the authority to select and remove a growing portion of the board of directors (ultimately the majority of the board) within 4 years, phasing in according to time- and funding-based milestones [Anthropic, 2023]. The Trust also has "protective provisions" requiring notice of actions that could significantly alter the corporation or its business. The structure is explicitly experimental, with "failsafe" provisions allowing changes through increasing supermajorities of stockholders as the Trust's power phases in. New Trustees are selected by existing Trustees, in consultation with Anthropic, and have no financial stake in Anthropic. The firm publicly announces new members [Anthropic, 2025]
OpenAI	**Update** In October 2025, OpenAI announced that it has completed its recapitalization. The nonprofit, now called the OpenAI Foundation, remains in control of the for-profit, and holds equity currently valued at approximately $130 billion. The recapitalization also grant the Foundation additional owernship as OpenAI's for-profit reaches a valuation milestone. The for-profit is now a public benefit corporation, called OpenAI Group PBC, which is required to advance its stated mission and consider the broader interests of all stakeholders, ensuring the company's mission and commercial success advance together. The OpenAI Foundation will initially focus on a $25B commitment across two areas: (1) Health and curing diseases (2) Technical solutions to AI resilience This builds on the $50M People-First AI Fund and the recommendations of the Nonprofit Commission. [OpenAI, 2025] It is also important to note that The Safety and Security Committee (SSC) will remain a committee of the OpenAI Foundation, and will continue its current role of providing governance over the safety and security practices of all of OpenAI, including OpenAI Group. [OpenAI] **Recap from Summer 2025** Uncommon governance structure. Founded as Non-profit as founders "initially believed a 501(c)(3) would be the most effective vehicle to direct the development of safe and broadly beneficial AGI while remaining unencumbered by profit incentives". Later incorporated a for-profit subsidiary (capped profit) to raise funds. For-profit controlled by non-profit and non profit legally bound to pursue the following mission of OpenAI: "To ensure that artificial general intelligence (AGI) benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome." For-profit arm has capped equity structure that limits maximum financial returns to investors and employees to balance profit incentives with safety concerns. Residual value will be returned to the Non-profit. The size of the cap is not transparent. Charter contains 'assist clause' to stop competing and assist a value-aligned, safety-conscious project to avoid race dynamics in late-stage AGI development [OpenAI] **Conversion plans:** In December 2024, OpenAI proposed a restructuring plan to convert the capped-profit into a Delaware-based public benefit corporation (PBC), and to release it from the control of the nonprofit. The nonprofit would sell its control and other assets, getting equity in return, and would use it to fund and pursue separate charitable projects. OpenAI's leadership described the change as necessary to secure additional investments. The plans provoked outside resistance and crisicsm. For example, a legal letter named "Not For Private Gain" [Not for Private Gain, 2025] asked the attorneys general of California and Delaware to intervene, stating that the restructuring is illegal and arguing how it would remove governance safeguards from the nonprofit and the attorneys general. In May 2025, the nonprofit's board chairman announced that the nonprofit would renounce plans to cede control after outside pressure. The capped-profit still plans to transition to a PBC, which critics said would diminish the nonprofit's control. [Fortune, 2025; CNBC, 2025; Reuters, 2025]
Google DeepMind	Same as AI Safety Index Summer 2025: For-profit company (part of Google)
Meta	Same as AI Safety Index Summer 2025: For-profit company
xAI	When xAI was incorporated in Nevada in March 2023, it was registered as a standard for-profit. It amended its corporate charter, turning into a benefit corporation in April 2023 , with the purpose "to create a material positive impact on society and the environment, taken as a whole." [1] However, in May 2024, xAI quietly amended its corporate charter again, terminating its status as a benefit corporation. After the status change, it has been representing itself in court still as a "Nevada benefit corporation" since November 2024, when it filed suit against OpenAI. It most recently claimed benefit corporation status to the court in May 2025, before the news that it changed its status went public. Nevada requires benefit corporations to report "all of its annual benefit reports, ... except that the compensation paid to directors and any financial or proprietary information included may be omitted." [LASST, 2025]
DeepSeek	Same as AI Safety Index Summer 2025: For-profit company
Z.ai	Same as AI Safety Index Summer 2025: For-profit company
Alibaba Cloud	For-profit company

Footnotes

[1] In the Summer 2025 edition, the stated purpose "to advance human scientific discovery and deepen understanding of the universe" was incorrectly attributed; this phrasing originated from a third-party article written by Grok, not from the company's own documentation. We hereby correct it.

Whistleblowing Protections

Indicator

Whistleblowing Policy Transparency

Definition

This indicator measures how fully and how accessibly an AI developer discloses its whistleblowing (WB) policy and system to the outside world. We look for a publicly reachable document (no paywall or login) that contains the material scope of reportable concerns, the people protected, the reporting channels offered (including anonymous options), oversight of the process, and the investigation and anti-retaliation guarantees. Evidence consists of artifacts that any external party can view, including public policy PDFs, dedicated "raise-a-concern" portals, relevant parts of safety frameworks, and transparency reports summarizing WB usage, outcomes, and effectiveness metrics.

Transparency Tiers:

2. No transparency
3. Fragments public: Parts of the design of the whistleblowing policy are public
4. Full policy public: Full policy, incl. processes, is public and highly transparent
    a. Full policy public + all details accessible: Policy does NOT refer to internal policies that are inaccessible to the public, but outside parties can fully review policy details (within reason)
    b. Effectiveness & Outcome transparency: The company provides details on the number of reports, topics, and follow-up actions, and also effectiveness, e.g., awareness & trust among employees, % of anonymous reports, appeal rates, whistleblower satisfaction, and types of cases received.

Why This Matters

Transparency on whistleblowing policies allows outsiders to assess the robustness of a firm's whistleblowing function. In AI safety contexts—where employees may be the first to spot concerning model behavior or negligent risk management—robust, visible policies are critical. Public posting subjects the company to scrutiny by regulators, journalists, and prospective staff for both the policy's quality and broader organizational culture around raising and addressing safety concerns. Private policies, on the other hand, can hide restrictive terms. Many large companies demonstrate high levels of transparency around internal whistleblowing systems (e.g., Microsoft, Volkswagen, Siemens), including by publishing annual whistleblowing statistics.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Anthropic doesn't have a public-facing whistleblower policy at the moment, but it plans to share more publicly in the near future, according to its company survey response. In Anthropic's Transparency Hub: Voluntary Commitments, the company identifies the three main channels through which its employees can report AI safety-related concerns. These mechanisms include confidentiality protections to ensure that employees can raise concerns without fear of retaliation. Anthropic has provided more information about its whistleblower policy in the company survey, including covered individuals, technical protections for confidentiality, protection for external reporting and anti-retaliation provisions. The quality of the whistleblower policy will be addressed in the indicator below.	**OpenAI has a public-facing whistleblowing policy ("OpenAI Raising Concerns Policy").** It includes aspects of covered violations, reporting mechanism (including Integrity Line), investigation mechanism for solutions, as well as confidentiality and no retaliation protection. OpenAI provided clarifications to its whistleblower policy in the company survey, including technical protections for confidentiality, anti-retaliation provisions, mechanisms to ensure effective investigation, and coverer concerns of the policy. The quality of the whistleblower policy will be addressed in the indicator below.	**Google DeepMind doesn't have a public-facing policy nor has not explained its reasons behind this decision, according to the Company Survey.** Google Code of Conduct delineates channels through which employees can raise their concerns towards different parties and the scope covered by such reporting. These concerns include a "no retaliation" clause. These measures apply to both employees and the extended workforce. Google shared more details about their whistleblowing policy in the company survey, including mechanisms to ensure effective investigation, investigation timeframes and procedures, confidentiality protection for internal and external reporting. The quality of the whistleblower policy will be addressed in the indicator below.	**Meta doesn't have a public-facing whistleblower policy at the moment, and it has not explained its reasons behind the decision publicly.** Its Code of Conduct referenced a Whistleblower and Complaint Policy, but it is not linked and not publicly retrievable. The Code delineates channels through which employees can raise their concerns, the mechanisms of investigation that follows, and "no retaliation" protections. Integrity line is available and linked, as well as harassment policy. The quality of the whistleblower policy will be addressed in the indicator below.	**xAI doesn't have a public-facing policy nor has not explained its reasons behind this decision, as according to the Company Survey.** However, xAI has stated that its employees have "whistleblower protections enabling them to raise concerns to relevant government agencies regarding imminent threats to public safety." Moreover, it has shared in the company survey more details, including the role designated to oversee the whistleblowing function, the investigative independence, the scope of policy, "no retaliation" and "confidentiality" protections towards employees, the reporting mechanisms etc. The quality of the whistleblower policy will be addressed in the indicator below.	**No public-facing whistleblower policy found.**	**No public-facing whistleblower policy found.** Z.ai skipped the whistleblower policy section in the company survey.	**No public-facing whistleblower policy found.** Its Code of Ethics states that employees have established whistleblower rules and procedures that are subject to update from time to time. The covered topics include violations of applicable laws or regulations, the Code, or Alibaba Group's related policies. Employees should report relevant information to the Compliance Officer. "No-retaliation" protection applies here.

Indicator

Whistleblowing Policy Quality Analysis

Definition

This analysis evaluates the quality of companies' whistleblowing policies based on all available evidence. The assessment analyzes 29 sub-indicators across five critical dimensions: 1) reporting channels and access, 2) whistleblower protections, 3) investigation processes, 4) system governance, and 5) AI-specific provisions.

Sub-indicators were derived from international reference standards—ISO 37002:2021, the ICC Guidelines, and the EU Whistleblowing Directive 2019/1937, which establish the gold standard for evaluation. Additional AI-specific items were included to address AI-specific concerns. For each Item, FLI evaluated the available evidence listed in the Whistleblowing Policy Transparency' indicator and rated the degree to which a company's policy satisfies it on a scale from 0 to 10, based on the publicly available information listed in the indicator on whistleblowing policy transparency and the company survey response, which includes whistleblowing policies, codes of conduct, safety frameworks, and survey responses.

Where no information was available, 0 points were assigned. The assessment measures how well firms' policies align with best practices while specifically examining whether companies have implemented specialized AI safety provisions, such as protections for reporting violations of safety frameworks.

Why This Matters

AI development's technical complexity and commercial pressures create unique risks that only insiders can identify, but safety culture needs to be prioritized. Robust whistleblowing policies with AI-specific protections serve as a critical last mile of defense when internal safeguards fail, enabling employees to report concerning behaviors, intentional deception, or capability discoveries that could pose catastrophic risks. Without robust protections, adequate coverage, and secure channels, companies can quietly abandon safety commitments while those best positioned to prevent harm remain silenced.

Title	Description	Anthropic	OpenAI	Google	Meta	xAI	DeepSeek	Z.ai	Alibaba
Overall average		4.7	4.8	4.9	2.2	1.0	0.0	0.0	0.5
Reporting Channels, Access, and Coverage		7	8	7	7	3	0	0	2
Protected Persons Coverage	Policy should at least cover current and former employees, contractors, shareholders, suppliers, former/prospective employees, and facilitators of reports	10	3	10	2	2	0	0	2
Policy Accessibility	Policy easily accessible to all covered persons	0	10	2	8	0	0	0	2
External Reporting Information & Rights	Policy must provide clear information about external reporting channels and right to approach these independently of internal processes, and explain or at least link to whistleblower protection rights	10	10	10	5	3	0	0	0
Multiple Reporting Channels	Offer multiple channels for reporting misconduct internally, incl. written, oral, in-person	9	9	9	9	9	0	0	2
Anonymous Two-Way Reporting	System enables fully anonymous reporting with secure two-way communication between reporter and investigators	10	10	10	10	5	0	0	0
Ombudsperson Channel	Reporting channel operated by an outsourced whistleblowing service provider.	0	10	0	10	0	0	0	0
Executive Oversight Channel	Separate reporting channel available for reports concerning senior executives (e.g. direct reporting line to board audit committee) or board members	7	5	10	5	0	0	0	0
Broad but clear material scope	Material scope covers at minimum potential violations of law, code of conduct. Ideally also further, broad categories, while retaining a high degree of clarity of what is in and out of scope.	8	8	8	5	7	0	0	7
Whistleblower Protections & Anti-Retaliation Measures		7	7	6	2	1	0	0	0
Confidentiality Protection	Strict protection required for reporter identity and any third parties mentioned in reports	10	10	2	8	0	0	0	0
Public Disclosure Protection	Protection for responsible media disclosure if internal and regulatory channels have failed or if there is an imminent or manifest danger to the public interest	10	10	10	0	0	0	0	0
List of Prohibited Practices and Anti-Retaliation Provisions	Policy must list comprehensive prohibited retaliatory actions with specific examples (demotion, harassment, termination, etc.), and explicit anti-retaliation provisions	10	10	10	0	0	0	0	2
Post-Investigation Monitoring	Active monitoring for retaliation continues for minimum 12 months after investigation concludes	0	0	0	0	0	0	0	0
NDA/Non-Disparagement Exceptions	Explicit statement that NDAs and non-disparagement agreements cannot prevent safety-related whistleblowing	10	7	7	0	0	0	0	0
Good Faith or Reasonable Cause Provisions	Clear good faith or reasonable cause standard that protects honest mistakes; high burden of proof required for false report sanctions	10	10	10	5	10	0	0	0
Handler/Investigator Protection	Explicit protections for employees who receive, investigate, or support whistleblowing reports	0	0	0	0	0	0	0	0

Table continues on next page

Title	Description	Anthropic	OpenAI	Google	Meta	xAI	DeepSeek	Z.ai	Alibaba
Investigation Process & Standards		1	2	3	1	0	0	0	0
Designated Impartial Receiver	Provably independent person or department must be designated to receive and handle reports - attached ideally to board	4	6	6	6	2	0	0	2
Seven-Day Acknowledgment	Written confirmation of report receipt must be provided within 7 days	0	0	0	0	0	0	0	0
Three-Month Feedback Timeline	Investigation status and follow up measures must be communicated to reporter within 3 months	0	0	2	0	0	0	0	0
Adequately Resourced Investigation Teams	Investigators must be independent from implicated departments and possess appropriate technical expertise for AI safety issues as well as sufficient resources to investigate effectively	0	5	5	0	0	0	0	0
Investigation Appeal Process	Formal right to appeal investigation outcomes to independent review body or board committee	0	0	0	0	0	0	0	0
System Governance & Quality Assurance		0	0	0	1	0	0	0	0
Comprehensive Effectiveness Metrics	Regular measurement tracking report outcomes, investigation timeliness, appeal rates, % of anonymous reports, retaliation incidents, and reporter satisfaction - not just volume	0	0	0	0	0	0	0	0
Data Retention and Deletion Policy	Clear policy specifying retention periods for reports and investigations (typically 5-7 years), secure deletion procedures, and data minimization principles	0	0	0	0	0	0	0	0
Secure Documentation System	Comprehensive audit trail with secure case management system and defined retention policies	0	0	0	5	0	0	0	0
Comprehensive Training Programs	Regular, role-specific training provided for all employees, specialized training for managers and investigators, ideally measuring training effectiveness.	0	0	0	0	0	0	0	0
Independent System Certification	Regular third-party audit and certification of whistleblowing system effectiveness and compliance	0	0	0	0	0	0	0	0
AI Safety-Specific Provisions		9	7	9	0	0	0	0	0
AI Safety Commitment Protection	Explicit protection for reporting violations of frontier safety frameworks (eg., RSP, Preparedness Frameworks), public AI safety commitments, and internal safety policies	10	10	10	0	0	0	0	0
AI Safety Coordination	Protection for AI risk reporting to dedicated AI safety bodies (UK AI Security Institutes, US Center for AI Standards and Innovation, or other international regulatory bodies)	10	10	10	0	0	0	0	0
AI risk transparency	Protections for reporting intentional deception of external evaluators, regulators or the public, suppression of publication of safety evaluation results, and inadequate disclosure of risk to regulators and the public,	8	0	8	0	0	0	0	0
Inadequate AI risk management and cybersecurity	Protections for reporting inadequate risk management processes, incl. assessment, monitoring, mitigation, deployment pressure despite concerning levels of risk, insufficient operational and cybersecurity practices incl. incidents	8	8	8	0	0	0	0	0

Indicator

Reporting Culture & Whistleblowing Track Record

Definition

This indicator evaluates whether an AI developer fosters a climate in which employees can raise safety-relevant concerns without fear of retaliation and with confidence that the concerns will be addressed. Evidence is drawn from (i) the organization's track-record of documented whistleblowing cases, (ii) the use, scope, and enforcement of non-disclosure or non-disparagement agreements (NDAs), (iii) leadership signals that encourage or discourage internal dissent, (iv) third-party evidence of psychological safety, and (v) patterns of safety information leaking externally (vi) departures linked to safety governance. The focus is on demonstrated behavior and outcomes rather than written policy statements. For whistleblowing incidents, we report individual names, concerns raised, and company response & status where available.

Notes of Best Practice: Companies should show a clear recent pattern of protecting and acting on employee safety reports; public commitment not to enforce legacy NDAs for safety topics; leadership statements praising internal critics; ≥ one anonymized psychological-safety survey with ≥ 70 % of staff agreeing "I can raise safety concerns without fear" and no credible retaliation cases in the last 24 months. Little public leaks as issues are addressed internally. Recent evidence (≤ 24 months) should be weighted twice as heavily as older cases to reward reforms.

Why This Matters

Whistleblowing policies can look impressive on paper, but they fail if the climate in the company suppresses reports, they're not effective when employees fear retaliation, or doubt anyone will act. This is why scrutinizing how firms respond to disclosures is critical. By focusing on actual cases, NDA practices, leadership signals, and exits tied to safety concerns, this indicator reveals which firms have built cultures where raising concerns feels like following protocol rather than betraying the company or colleagues—the trust and accountability needed for early detection of catastrophic AI risks.

Table begins on the next page

EU AI Code of Practice Safety and Security	**Measure 8.3** Examples of a healthy risk culture include annually informing workers of the Signatory's whistleblower protection policy and making such policy readily available to workers such as by publishing it on their website.
Anthropic	Summer 2025 Index highlighted Anthropic's public renouncement of the use of non-disparagement clauses in severance agreements (July 2024) Since the Summer 2025 iteration, there has been no known whistleblower or retaliation incidents publicly reported. In September 2025, the company publicly endorsed California's SB 53, which explicitly includes requirements for whistleblower protections to reports of violations of the bill's requirements as well as disclosures of specific, substantial dangers to public health or safety.
OpenAI	Summer 2025 Index highlighted that OpenAI's internal culture has been marked by safety-driven resignations and public disputes over non-disparagement and equity-clawback clauses, culminating in a June 2024 "Right-to-Warn" movement calling for stronger whistleblower rights. Since the Summer 2025 iteration, there has been no known whistleblower or retaliation incidents publicly reported.
Google DeepMind	Summer 2025 Index highlighted Google's record of repeated conflicts between management and employees raising ethical or scientific objections, with several high-profile dismissals often framed by the company as security or academic disputes. Since the Summer 2025 iteration, there has been a new whistleblower case: **William Huesman (November, 2025):** The former Google Cloud director said he resigned from his position in February 2024 after his supervisor "undermined, marginalized and ultimately blacklisted" him, according to his complaint filed in November 2025 in the US District Court for the Middle District of Florida. He claimed that the retaliation came as a result after he reported the repeated misconduct—including frequent intoxication at work and over 20 HR complaints of his supervisor, Snehanshu Shah, a Managing Director at Google. Google hasn't responded to a request for comment. [Bloomberg Law, 2025] [Human Resources Director, 2025]
Meta	Summer 2025 Index highlighted that Meta has faced multiple legal and reputational challenges for suppressing internal dissent through overbroad non-disparagement and confidentiality clauses later ruled illegal by the NLRB. Since the Summer 2025 iteration, there has been updates on **Sarah Wynn-Williams**' case (former director of global public policy at Meta's precursor, Facebook) Louise Haigh, a UK Member of Parliament, publicly accused Meta of trying to "silence and punish" Wynn-Williams, and said that Wynn-Williams was "facing a fine of $50,000 every time she breached an order secured by Meta preventing her from talking disparagingly about the company." Meta defends that, since she voluntarily signed the non-disparagement agreement, so she must abide by it [Guardian, 2025]
xAI	**Project Skippy leak:** In July 2025, Internal documents and Slack messages from xAI leaked to Business Insider revealing an internal project called "Project Skippy," which asked more than 200 employees to record videos of their own faces and conversations to train Grok to recognize human emotions and expressions. The disclosure, made by anonymous insiders concerned about potential misuse of their likenesses and consent forms granting xAI "perpetual" rights to their biometric data, functioned as a semi-whistleblower leak highlighting employee unease over privacy and data ethics. As of late 2025, neither Elon Musk nor xAI has issued any public response or clarification regarding the project or the concerns raised. [Business Insider, 2025]
DeepSeek	No public or media record of reported whistleblower or retaliation incidents, NDA disputes or changes, leaks of internal information.
Z.ai	No public or media record of reported whistleblower or retaliation incidents, NDA disputes or changes, leaks of internal information.
Alibaba Cloud	**Sexual Assault Whistleblower (Ms.Zhou)**: In August 2021, an Alibaba employee publicly accused her manager and a client of sexual assault after internal complaints were ignored. Her post went viral on Alibaba's intranet and Chinese social media, forcing the company to act. Alibaba fired the accused manager but later terminated the whistleblower herself in November 2021, citing "spreading false information" and "damaging the company's reputation," as well as dismissing 10 other employees that publicized the event internally. Daniel Zhang, who is the CEO at the time, condemned the incident as "shameful" and promised zero tolerance for harassment, but did not respond to the retaliation of the whistleblower herself. In December 2021, Alibaba executive Li Yonghe — a vice president who resigned over the scandal — filed a defamation lawsuit against the employee, alleging that her public accusations had damaged his reputation, and claiming that he had not ignored Zhou's complaint. [Guardian, 2021; DW, 2021]

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Governance and Accountability

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Clear, enforceable accountability across all levels; strong whistleblowing, legal, and oversight systems.
- **B** Defined governance roles and accountability measures; minor gaps in enforcement or transparency.
- **C** Basic accountability mechanisms; limited clarity or inconsistent application.
- **D** Weak governance; vague roles and limited channels for reporting or oversight.
- **F** No credible accountability framework; governance absent or nominal.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Domain

Information Sharing & Public Messaging

This domain evaluates how openly companies share technical, safety, and governance information, and how their public and legislative messaging align with responsible AI governance.

Table of Contents

Chinese Regulatory System Summary

Mandatory reporting under the Interim Measures requires AI providers to remove unlawful content, retrain affected models, and notify authorities.

The AI Safety Governance Framework 2.0 functions as non-binding policy guidance, encouraging broader risk and vulnerability information sharing, database establishment, and international cooperation to address systemic and cross-border AI safety risks.

National Binding Instruments

Serious Incident Reporting & Government Notifications

Interim Measures (Article 14) requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. Deep-Synthesis Provisions (Jan 2023) Service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments

Strategic and Policy Guidance Documents

The AI Safety Governance Framework 2.0

Article 5.9 emphasizes sharing information on AI safety risks and threats, which requires tracking and analyzing security vulnerabilities, defects, risk threats, and security incidents related to AI technologies, products, and services. The clause calls for the establishment of an AI vulnerability information database and a risk and threat information-sharing mechanism that covers developers, service providers, and professional technical institutions. It also encourages international exchange and cooperation in AI safety risk and threat information-sharing, calling for the development of relevant cooperation mechanisms and technical standards to jointly prevent and respond to large-scale, cross-domain diffusion of AI safety risks.

Technical Specifications

Indicator

System Prompt Transparency

Definition

This indicator evaluates how openly companies disclose the instructions—known as system prompts—that guide how their most advanced AI systems behave. These prompts define an AI system's behavior and safety performance. Full transparency involves releasing the exact prompts used in deployed systems, keeping version histories, and explaining how and why key design decisions were made. Relevant evidence may be collected from model documentation, technical reports, or transparency pages.

Why This Matters

System prompts directly control how an AI system interprets and filters user inputs, and therefore undisclosed prompts make it difficult for outside experts to verify safety claims or replicate results. Publishing them enables independent analysis of whether built-in safeguards work as intended and shows a company's willingness to subject its implementation choices to public and scientific scrutiny.

EU AI Code of Practice Safety and Security		**Measure 7.1** Signatories will provide in the Model Report a specification of how Signatories intend the model to operate (often known as a "model specification"), including by: (a) specifying the principles that the model is intended to follow; (b) stating how the model is intended to prioritise different kinds of principles and instructions; (c) listing topics on which the model is intended to refuse instructions; and (d) providing the system prompt.
Anthropic	Claude	**Update** Shared prompts: (and # of updates) Haiku 4.5 (1) Sonnet 4.5 (1) Opus 4.1 (1) Opus 4 (3) Sonnet 4 (3) **Recap from Summer 2025** Since August 2024, Anthropic publicly shares the system prompts for the Claude.ai web interface and mobile apps. They further committed to log changes they make to these prompts online. These system prompt updates do NOT apply to the Anthropic API. Shared prompts: (and # of updates) Opus 4 (1) Sonnet 4 (1) Sonnet 3.7 (1) Sonnet 3.5 (4) Opus 3 (1) Claude Haiku 3 (1) Simon Willison reported that the publicly shared version does not include the description of various tools available to the model [Simon Willison, 2025].
OpenAI	ChatGPT	No transparency on system prompts for frontier systems.
Google DeepMind	Gemini	No transparency on system prompts for frontier systems.
Meta	Llama	No transparency on system prompts for frontier systems.
xAI	Grok	**Update** Through its public Github repository, the company regularly releases the full text of the system prompt used across its Grok product suite. It is openly available for inspection and reuse under the GNU Affero General Public License v3.0, which The repository currently includes prompts for (1) Grok 4 on grok.com and X (2) Grok 3 (3) Grok Explain feature on X (4) Grok bot on X (5) injected prefix prompts for API-served Grok models. **Recap from Summer 2025** After two incidents involving unauthorized system prompt changes—one in February 2024 causing political censorship and another in May 2025 leading Grok to make racially charged statements—xAI responded by publicly releasing its Grok system prompts on GitHub and committing to keep them regularly updated.
DeepSeek	R1	No transparency on system prompts for frontier systems.
Z.ai	GLM-4.6	No transparency on system prompts for frontier systems.
Alibaba Cloud	Qwen3-Max	No transparency on system prompts for frontier systems.

Indicator

Behavior Specification Transparency

Definition

This indicator assesses whether companies publish detailed specifications outlining their models' intended behaviors, boundaries, and decision-making frameworks. For companies that shared such documents, we provide high-level summaries and link to the sources. We include documents that concretely outline the goals, values, and behavioral guidelines that developers aim to instill in their models. Documentation should explain how developers want their models to handle various scenarios, conflicts, and edge cases, and detail how these values are implemented, including metrics or evidence of how well these values are achieved in practice. Specifications should ideally be current and include a tracked version history with dates. Important aspects are specificity, comprehensiveness across use cases, and inclusion of concrete examples. Internal training documents, vague mission statements, and brief high-level descriptions are not in scope.

Why This Matters

Behavioral specifications clarify what companies intend their AI systems to do, offering a higher-level view of safety and value alignment than technical prompts alone. Publishing these specs enables external verification of whether deployed models match stated intentions and allows identification of gaps in safety considerations. Companies willing to specify and publish concrete behavioral guidelines demonstrate accountability for their choices and enable public scrutiny.

EU AI Code of Practice Safety and Security		**Measure 7.1** Signatories will provide in the Model Report a specification of how Signatories intend the model to operate (often known as a "model specification"), including by: (a) specifying the principles that the model is intended to follow; (b) stating how the model is intended to prioritise different kinds of principles and instructions; (c) listing topics on which the model is intended to refuse instructions; and (d) providing the system prompt.	
Anthropic	Claude	**Update** Sonnet 4.5 system card does not refer to Constitutional AI, and instead emphasizes reinforcement learning from human feedback and from AI feedback as the main post-training technique. The document's alignment and safety sections discuss evaluation awareness, automated behavioral audits, interpretability studies, and responsible scaling safeguards, but none describe a normative ruleset guiding model behavior. **Recap from Summer 2025** Constitutional AI: Method for training AI systems to be harmless by using a set of written principles (a "constitution") rather than relying solely on large-scale human feedback. **What it's for:** 1) Supervised learning phase: Model self-critiques and revises its outputs based on constitutional principles, creating a supervised learning dataset 2) RLAIF phase: Model compares response pairs using constitutional principles to generate preference labels, then trains via RL on these AI-generated preferences **Timeline & Development:** December 2022: Original Constitutional AI paper published May 2023: Claude's constitution made public (58 principles)	**Constitution (May 2023):** 58 principles (1.2k word) drawn from: - UN Declaration of Human Rights - Apple's Terms of Service - DeepMind's Sparrow principles - Non-Western perspectives - Anthropic's own research Example principle: "Please choose the response that most supports and encourages freedom, equality, and a sense of brotherhood." **Limitations:** (1) Version uncertainty: Only May 2023 constitution is public; current production versions unknown (2) Attribution ambiguity: Anthropic reports using multiple post-training techniques—human feedback, Constitutional AI, and the modeling of specific character traits—making it unclear how much influence any single method exerts on final model behavior. (3) Transparency gap: No public commitment to sharing constitution updates. (4) Behavioral indeterminacy: Since the AI itself determines how to balance competing constitutional principles, Anthropic's approach does not explicitly specify the intended behavior of its AI systems, especially when values conflict.

Table continues on next page

OpenAI	ChatGPT	**Update** The latest Model Spec update (Oct, 2025) introduces three main changes (1) **Expanding guidance on mental health and well being** in the self-harm section, covering delusional and manic behavior, with concrete examples for the models to behave with empathy and grounding (2) **New section on "respect real-world ties,"** instructing models to support users' real-world relationships and discourage dependence on the AI assistant, particularly in contexts involving loneliness, emotional intimacy, or personal advice (3) **Clarification on "chain of command" delegation**, specifying that the models can treat outputs from tools as authoritative when doing so matches user intent and prevents errors or confusion **Recap from Summer 2025** **OpenAI Model Spec** OpenAI's Model Spec is a detailed (~28k words), public, living rule-book that defines the objectives, safety rules, and default behaviours OpenAI trains its models —via human feedback and deliberative alignment—to follow. **What it's for** 1) Human RLHF guidance – provides a single, public rule-book labelers follow when creating preference data. 2) Deliberative Alignment – o-series models (o1, o3, o4-mini) are explicitly taught to read and reason over the Spec before answering. 3) Automated evaluation – OpenAI ships a challenge-prompt suite to measure adherence. **Timeline & Versions** 1st May 2024 2nd Feb 2025 3rd Apr 2025	**Framework** Three principle types 1) Objectives – broad goals such as "assist the developer & end user" and "benefit humanity." 2) Rules – hard, platform-level constraints (e.g. comply with law, prohibit or restrict certain content, protect privacy, uphold fairness). 3) Defaults – stylistic and behavioural norms that developers/users may override. Sections: - Stay in bounds - Seek the truth together - Do the best work - Be approachable - Use appropriate style. Includes specific guidance on specific policy areas such as poticial, medical, or harmful content. Risk taxonomy: - Misaligned goals - Execution errors - Harmful instructions. Chain of command: Platform (OpenAI) → Developer → User → Guideline → Untrusted text. Within any level, explicit > implicit, later > earlier. (OpenAI's Usage Policy overrides the Spec if the two conflict.) Ongoing Development: Released under CC0 license (public domain) Changelog and version history maintained on GitHub OpenAI commits to regular updates as the spec evolves
Google DeepMind	Gemini	No detailed specification available	
Meta	Llama	No detailed specification available	
xAI	Grok	No detailed specification available	
DeepSeek	R1	No detailed specification available	
Z.ai	GLM-4.6	No detailed specification available	
Alibaba Cloud	Qwen3-Max	No detailed specification available	

Voluntary Commitment

Indicator

G7 Hiroshima AI Process Reporting

Definition

The G7 Hiroshima AI Process (HAIP) Reporting Framework is a voluntary transparency mechanism launched in February 2025 for organizations developing advanced AI systems. Organizations complete a comprehensive questionnaire covering seven areas of AI safety and governance practices, including risk assessment, security measures, transparency reporting, and incident management. All submissions are published in full on the OECD transparency platform. This indicator tracks whether firms participated in HAIP as a measure of their commitment to AI safety transparency

Why This Matters

The HAIP framework represents the first globally standardized mechanism for AI developers to disclose their safety practices in comparable detail. Participation creates reputational stakes and enables external scrutiny since reports are published. Organizations choosing to participate signal a willingness to be held accountable and contribute to collective learning.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	z.AI	Alibaba Cloud
Substantive Submission [OECD, 2025]	Substantive Submission [OECD, 2025]	Substantive Submission [OECD, 2025]	No Submission	No Submission	No Submission (Not based in G7 nation)	No Submission (Not based in G7 nation)	No Submission (Not based in G7 nation)

Indicator

EU General-Purpose AI Code of Practice

Definition

The AI Act Code of Practice (introduced in EU AI Act Article 56) is a set of guidelines for compliance with the AI Act. It is a crucial tool for ensuring compliance with the EU AI Act obligations, especially in the interim period between when General Purpose AI (GPAI) model provider obligations came into effect (August 2025) and the adoption of standards (August 2027 or later). Though they are not legally binding, GPAI model providers can adhere to the Code of Practice to demonstrate compliance with GPAI model provider obligations until European standards come into effect. [EU AI Act, 2025]

Why it matters

AI companies' participation demonstrates its readiness to meet forthcoming regulatory obligations and willingness to align with the EU's risk-based approach.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	z.AI	Alibaba Cloud
Signed	Signed	Signed	Declined to sign	Signed up to the Safety and Security Chapter	No public stance	No public stance	No public stance

Indicator

Frontier AI Safety Commitments (AI Seoul Summit, 2024)

Definition

Announced at the AI Seoul Summit in May 2024, the Frontier AI Safety Commitments are voluntary pledges by leading AI developers to aim for safe and responsible development and deployment of highly capable general-purpose AI systems. [UK Department for Science, Innovation and Technology, 2025] An important component of the Commitment is that companies have agreed to publish a safety framework intended to evaluate and manage severe AI risks.This directly correlates with the "Safety Framework" section of the Index.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	z.AI	Alibaba Cloud
Signed The safety framework is published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational, as according to the company survey response.	**Signed** Safety Framework is published and implementation in progress.	**Signed** The safety framework is published, although the extent of implementation is not clear.	**Signed** The safety framework is published, although the extent of implementation is not clear.	**Signed** The safety framework is published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational, as according to the company survey response.	**Not Signed**	**Signed** Safety Framework is published & Implementation in progress. According to their company survey response, safety Framework is published & Implementation in progress. However, no public framework is found online.	**Not Signed**

Indicator

FLI AI Safety Index Survey Engagement

Definition

We report which companies have engaged with our index survey to voluntarily disclose additional information. Full survey responses are linked below.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	z.AI	Alibaba Cloud
Survey Response Submitted [Company Survey]	Survey Response Submitted [Company Survey]	Survey Response Submitted [Company Survey]	None Received	Survey Response Submitted [Company Survey]	None Received	Survey Response Submitted [Company Survey]	None Received

Indicator

Endorsement of the Oct. 2025 Superintelligence Statement

Definition

The October 2025 Superintelligence Statement is an open letter, endorsed by a broad coalition of policy makers from all sides, industry, faith leaders, and researchers etc. The letter calls for a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in.

Why it matters

Endorsement matters because it publicly commits organizations and individuals to restraint at the highest capability frontier, reinforcing precautionary governance norms and prioritizing global safety over competitive acceleration, especially if such endorsement comes from the leadership level.

Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	z.AI	Alibaba Cloud
6 current staff members have signed, but nobody from the corporate leadership	4 current staff members have signed, but nobody from the corporate leadership	5 current staff members have signed, but nobody from the corporate leadership	None	None	None	CEO Peng Zhang has signed the statement.	None

Risks & Incidents

Indicator

Serious Incident Reporting & Government Notifications

Definition

This indicator evaluates incident reporting commitments, frameworks, and track records. For frameworks and commitments, the indicator assesses whether companies have publicly discussed any systems and commitments to share critical information about red-line incidents or capabilities with government bodies (e.g., US CAISI, UK AISI), peer organizations, or the public. Such incidents can include successful large-scale misuse, near-miss events, scheming by AI models, and identified model capabilities with severe national security implications. The indicator further tracks relevant incident documentations that the company has already shared. Evidence comes from safety frameworks, documented reporting procedures, participation in information-sharing agreements, and public incident reports.

Notes on Best Practice: Clear public commitments to report specific categories of incidents to government bodies, with documented procedures for incident classification and escalation. Information-sharing agreements with disclosed scope, publishing reports on recent incidents, demonstrating transparency about warning signs discovered during development, and establishing clear thresholds for mandatory reporting, specificity, and comprehensiveness of reporting commitments.

Why This Matters

Proactive incident reporting enables collective learning from safety failures and near-misses across the AI industry, preventing repeated mistakes and identifying emerging risks before they materialize. Transparency about dangerous capabilities and misalignment incidents is critical for government oversight. Without such transparency, companies may make deployment decisions based on marginal safety improvements while baseline risks remain unacceptably high.

EU AI Code of Practice Safety and Security	**Commitment 9 (Measure 9.1-9.4)** Signatories are required to adopt additional measures to track serious incidents, including monitoring external sources such as media reports, research papers, and incident databases, and enabling downstream developers, users, and third parties to report incidents through clear channels. When reporting to relevant authorities, signatories must include details such as the incident timeline, harm caused, affected parties, chain of events, model involvement, corrective actions, and root cause analysis. Reporting must occur promptly—within 2 to 15 days depending on the severity of the incident—followed by updates every 4 weeks until resolution and a final report within 60 days after resolution. All related documentation must be retained for at least five years.
Anthropic	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** Responsible Scaling Policy contains a broad voluntary commitment on ASL disclosing ASL levels: - "We will notify a relevant U.S. Government entity if a model requires stronger protections than the ASL-2 Standard" **Public transparency reports:** Anthropic has regularly published comprehensive misuse reports which documents real-world cases of actors attempting to exploit Claude for malicious purposes, along with detection methods and enforcement actions taken. - August 2025 - "Threat Intelligence Report: August 2025" - March 2025 – "Misuse Monitoring and Response Report" Other: - Platform Security Transparency Hub provides some enforcement statistics including #banned accounts for Usage Policy violations, number of appeals processed, CSAM reports to NCMEC, and law enforcement requests. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum2025].

OpenAI	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** Regular reports documenting their disruption of malicious uses of their AI systems. Comprehensive reports detail enforcement actions against state-affiliated threat actors and covert influence operations identify specific threat groups (e.g., Storm-2035, Spamouflage), quantify disruptions (accounts banned, operations terminated), and describe the tactics employed (phishing, malware development, influence campaigns, election interference). - Feb 2024 – "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors" - May 2024 – "Disrupting a Covert Iranian Influence Operation" - Jun 2024 – "Update on Disrupting Deceptive Uses of AI" - Aug, 2024: "Disrupting a covert Iranian influence operation" - Oct 2024 – "Influence and cyber operations: an update" - Feb 2025 - "Disrupting malicious uses of our models" - Jun 2025 - "Disrupting malicious uses of AI" - Oct 2025 - "Disrupting malicious uses of AI" **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025]. Comments on incident response from index survey (Q31) [Response]: "OpenAI has developed and continues to improve incident response programs across key areas of its operations, and is likewise improving and iterating on AI safety incident-specific protocols that are tailored to our operations and technology. Our goal is to respond to incidents in a rapid, coordinated way. [...] Incident Response Capabilities include (1) **Technical Controls for Rapid Mitigation:** We maintain the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized within the last year in response to our finding that a GPT-4o model update was overly flattering or agreeable (see Sycophancy in GPT-4o: what happened and what we're doing about it) (2) **Incident Response Planning and Structure:** OpenAI has formal incident response plans for key areas of operations, including AI safety incident-specific protocols. Our response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. We maintain close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans.
Google DeepMind	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** Frontier Safety Framework 3.0 states that "If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate safety of frontier AI," a commitment it has kept from the last version of the Frontier Safety Framework 2.0. [Google, 2025]. **Public transparency reports:** Relevant publications: - 'Adversarial Misuse of Generative AI' (January 2025) - Detailed how threat actors—from scammers to state-aligned groups—attempt to misuse Google Gemini in deception, persuasion, and cyber operations. Described mitigation strategies and detection tooling [Google ,2025]. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025].

Table continues on next page

Meta	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** Meta consistently issues quarterly integrity reports about its platforms [Meta, 2024], these include reports on disrupting adversarial threat such as influence operations [Meta, 2025]. No reports for frontier AI models available. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025].
xAI	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** xAI mentions in its RMF that it aims for "public transparency" about its risk management policies and intends to publish updates but has not mentioned whether it is going to publish misuse and model misalignment report. **Industry information sharing:** There is no publicly visible evidence that xAI systematically shares incident-data or model-failure information with industry partners.
DeepSeek	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.
Z.ai	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments Its survey response (Q31) has indicated that the company has implemented the following capability: (1) Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). (2) Successfully tested rapid full model rollback including internal deployments within the last 12 months. (3) Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. (4) Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. (5) Conducted at least one full live emergency response drill/simulation in the past 12 months. (6) Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander, and a 24 × 7 duty roster. (7) Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.
Alibaba Cloud	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.

Table continues on next page

Indicator

Extreme-Risk Transparency & Engagement

Definition

The indicator assesses the extent to which companies and their leadership (A) publicly recognize the potential for catastrophic AI harm and (B) proactively communicate about them in an evidence-based and analytically grounded manner. The criteria are frequency, specificity, and prominence of communication about AI's potential for catastrophic outcomes (including existential risks, mass casualties, or societal-scale disruption).

Evidence includes official blogs, testimonies, leadership communications, including signed statements. Excludes technical safety papers, model cards, and formal safety frameworks (captured in separate indicators).

Why This Matters

Public communication about AI's potential for catastrophic outcomes shapes societal preparedness, policy responses, and research priorities. Companies developing frontier AI possess unmatched knowledge of actual capabilities, near-term developments, and observed warning signs. Their leadership's willingness to transparently discuss extreme risks indicates a precautionary approach and enables an informed discourse on policy and national security.

Anthropic	**Update** CEO Dario Amodei released a statement on Anthropic's commitment to American AI leadership, where he emphasizes that Anthropic was founded on the principle that AI should advance "human progress, not peril," which means that "making products that are genuinely useful, speaking honestly about risks and benefits, and working with anyone serious about getting this right." [October2025] **Recap from Summer 2025** Company communication and its leaders regularly and pro-actively communicate extreme risks. Anthropic CEO Dario Amodei's quotes in the past: - Warns AI may eliminate 50% of entry-level white-collar jobs within the next five years [Business Insider, 2025] and says on television that he is "raising the alarm" about this [CNN, 2025]. - Blog post calling the Paris AI Action summit a "missed opportunity", saying ".. greater focus and urgency is needed on several topics given the pace at which the technology is progressing." [Anthropic, 2025]. - Warned Congress that AI could enable bioweapon creation within 2-3 years [Bloomberg, 2023]. - Repeatedly warns that 'powerful AI', which he likens to "a country of geniuses in a datacenter", could arrive as early as 2026 or 2027, and is explicit about extreme risks [Anthropic, 2025]: ".. hardcore misuse in AI autonomy that could be threats to the lives of millions of people. That is what Anthropic is mostly worried about." [Business Insider, 2025] CAIS statement on extinction risk signed by: Dario Amodei (CEO), Daniela Amodei (President), Jared Kaplan (co-founder), Chris Olah (co-founder)
OpenAI	**Update** The company released an update discussing AI progress and recommendations (November, 2025). It includes a discussion of AI safety and superintelligence safety, quoted as below: "OpenAI is deeply committed to safety, which we think of as the practice of enabling AI's positive impacts by mitigating the negative ones. Although the potential upsides are enormous, we treat the risks of superintelligent systems as potentially catastrophic and believe that empirically studying safety and alignment can help global decisions, like whether the whole field should slow development to more carefully study these systems as we get closer to systems capable of recursive self-improvement. Obviously, no one should deploy superintelligent systems without being able to robustly align and control them, and this requires more technical work." CEO Sam Altman appears to have tempered his warnings: from early concerns about "lights out for all of us" [Business Insider, 2023] and "human extinction", his 2025 post "the Gentle Singularity" suggests that "living through [the singularity] will feel impressive but manageable" partly because "society is resilient, creative, and adapts quickly" **Recap from Summer 2025** Corporate communication and its leadership sometimes talk about extreme risks. CEO Altman's communications have changed over time and become slightly more optimistic. OpenAI CEO Sam Altman's quotes in the past: - In 2015, he stated: "I think that AI will probably, most likely, sort of lead to the end of the world" [Standford, 2024], and published a blog on "why machine intelligence is something we should be afraid of" [Altman, 2015]. - In 2023, he published a blog "Planning for AGI and Beyond," stating OpenAI will proceed as if risks are "existential" [OpenAI, 2023]. - In another blog, argued about the need for global coordination on the governance of superintelligence, and that "it would be important that such an agency focus on reducing existential risk" [OpenAI, 2023]. - In his 2023 Senate testimony, he urged lawmakers to implement federal licensing and external audits to bound risk [Time, 2023]. - In his recent communications, Altman adopted a more optimistic tone. In his recent congressional testimony, Altman told lawmakers that requiring government approval would be "disastrous" for US AI leadership [Washington Post, 2025]. CAIS statement on extinction risk signed by Sam Altman (CEO), Adam D'Angelo (board member), Wojciech Zaremba (cofounder)

Table continues on next page

Google DeepMind	**Update** Google DeepMind has updated its Frontier Safety Framework twice in 2025, February and September, respectively, taking into consideration more frontier and extreme AI risks such as "rogue AI." **Recap from Summer 2025** Corporate communications rarely mention extreme risks. Google Deepmind's leadership regularly discusses extreme risks in media interviews. Google's leadership does not. Quotes in the media from leadership: Demis Hassabis (CEO) - "We must take the risks of AI as seriously as other major global challenges, like climate change [...] It took the international community too long to coordinate an effective global response [..]. We can't afford the same delay with AI" [Guardian, 2024]. - "Artificial intelligence is a dual-use technology like nuclear energy: it can be used for good, but it could also be terribly destructive" [Time, 2025]. - Demis shares that he thinks AGI is only a "handful of years away" and that he is very worried about deception, calling it "incredibly dangerous", and speaks about encouraging the Security institutes to investigate them [Youtube, 2025]. Other examples: [CNN, 2025] [CBS, 2025] Shane Legg (Chief AGI Scientist) communicates a similar stance, and he recently stated AI is a very powerful technology, and it can and should be regulated." [Axios,2025]. In contrast, at the 2025 AI Action Summit in Paris, Google's CEO Sundar Pichai stated that "The biggest risk could be missing out." [Observer, 2024] CAIS statement on extinction risk signed by Demis Hassabis (CEO), Shane Legg (Co-Founder), Lila Ibrahim (COO).
Meta	**Update** Company and leadership rarely address extreme risks. **Recap from Summer 2025** Mark Zuckerberg and Chief AI Scientist Yann LeCun express the strongest counter narrative to AI existential risk concerns among major companies [Interesting Engineering, 2025]. LeCun does not believe that AI poses existential risk and calls such concerns "complete B.S.", arguing we need "the beginning of a hint of a design for a system smarter than a house cat before worrying about superintelligence" [Tech crunch, 2024]. Meta's president of global affairs expresses a similar position [Politico, 2024], comparing the discussion and framing the topic as a "moral panic" [Independent, 2024]. Zuckerberg is concerned about power concentration: "But I stay up at night worrying more about an untrustworthy actor having the super strong AI, whether it's an adversarial government or an untrustworthy company or whatever.". He shares that:" Bioweapons are one of the areas where the people who are most worried about this stuff are focused, and I think it makes a lot of sense.". He expresses less urgency on existential risk addressing deception as "longer-term theoretical risks", and saying ". we focus more on the types of risks that we see today .." [Dwarkesch Podcast, 2024].
xAI	**Update** **Corporate communication itself does not publicly share information about extreme risks. CEO Musk has a track-record of raising concerns.** Elon Musk argued that "Long-term, AI's gonna be in charge, to be totally frank, not humans. If artificial intelligence vastly exceeds the sum of human intelligence, it is difficult to imagine any humans would actually be in charge" [Pravda, 2025] **Recap from Summer 2025** In 2014, Musk called AI humanity's "biggest existential threat.", calling for regulatory oversight [Live Science, 2014]. In September 2023, he told senators "'there's some chance – above zero – that AI will kill us all." [NBC, 2023]. At the 2024 Saudi summit, he estimated a "10-20% chance AI goes bad." [Fortune, 2025]. CAIS statement on AI Risk signed by: Igor Babuschkin (co-founder), Tony Wu (co-founder)
DeepSeek	**Update** **Researchers and policy teams at the companies are increasingly engaging with the topic of existential risks, signaling a more public engagement of the company in the field.** DeepSeek researcher Chen Deli struck a conspicuously pessimistic note about the future of AI at a major state-backed tech conference on Friday, warning about its potentially "dangerous" impacts on society and the job market. Chen said he was optimistic about the tech itself but pessimistic about its overall impact on society: "Humans will be completely freed from work in the end, which might sound good but will actually shake society to its core." In September, 2025, DeepSeek's head of AI governance spoke at an open-source conference about ethical guardrails. [SCMP, 2025] **Recap from Summer 2025** The company and its leadership do not discuss extreme risks from AI. CEO Liang Wenfeng keeps a very low profile and rarely speaks in public. Beijing instructed DeepSeek "not to engage with the media without approval." [Reuters, 2025].
Z.ai	**Update** **Z.ai Corporate communications don't speak about the potential for extreme risks. Leadership has been more actively engaging with the subject.** In October 2025, Z.ai's CEO Peng Zhang signed the FLI's superintelligence statement, calling for a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in. **Recap from Summer 2025** While corporate communication rarely discusses catastrophic and existential risks, the company's Chief Scientist Tang Jie and its CEO has acknowledged the need to get prepared for existential risks and align super intelligent systems.
Alibaba Cloud	**Corporate communications and the company's leadership rarely engage with the subject publicly.**

Public Policy

Indicator

Policy Engagement on AI Safety Regulations

Definition

This indicator tracks a company's involvement in proactively shaping or responding to laws and regulations concerning AI safety. Evidence includes public statements, consultation submissions, testimony, and official responses, participation in trade associations or coalitions that lobby on safety-related issues, as well as active participation in drafting relevant regulations and standards.

Why it matters

Leading AI developers have unique technical expertise and credibility to advise governments on charting a responsible path for this transformative technology. Tracking patterns in companies' engagements on specific regulations can indicate which firms take a proactive stance on raising the bar for sensible protections.

Anthropic	**Update** **California SB 53** SB 53 provides "a blueprint for evidence-generating transparency measures" for governing frontier AI systems. [Carnegie Endowment, 2025] Anthropic publicly endorsed SB 53, calling it a "trust-but-verify" approach that strengthens accountability for frontier AI systems and sets a strong baseline for transparency. The company emphasized that while it still prefers a federal framework, California's action is necessary given the rapid development of advanced models [Anthropic, 2025; TechCrunch, 2025] **Preemption of state-level AI legislation** In its endorsement announcement for California SB 53, it stated that "frontier AI safety is best addressed at the federal level instead of a patchwork of state regulations," deviating from its previous stance on state-oriented AI safety approach.	**Recap from Summer 2025** **EU AI Act** N/A **US Legislations** California SB 1047 Anthropic raised initial concerns about key provisions, but the CEO later expressed cautious support, acknowledging that the benefits of the bill likely outweight its costs. It also actively shape the final version of the legislation. New York Raise Act N/A Preemption of state-level AI legislation In 2025, Anthropic opposed federal efforts to preempt state-level AI laws. CEO Dario Amodei argued that states should retain authority to set transparency and safety standards, warning that federal preemption could weaken oversight.
OpenAI	**Update** **California SB 53** No public stance **Preemption of state-level AI legislation** In OpenAI's letter to Governor Newsom on harmonized regulation, the company urges California to "harmonize" with federal and global frameworks instead of layering its own additional requirements. OpenAI argues that "a patchwork of state rules… could slow innovation without improving safety," urging California instead to align with "federal and global safety guidelines" to "avoid duplication and inconsistencies between state requirements and the safety frameworks already being advanced by the US government and our democratic allies."	**Recap from Summer 2025** **EU AI Act** In 2023, OpenAI lobbied EU officials to weaken parts of the AI Act, arguing that foundation models such as GPT-4 should not face strict obligations unless adapted for specific uses. **US Legislations** California SB 1047 In 2024, OpenAI opposed California's SB 1047, arguing that its safety requirements—such as third-party evaluations and incident reporting—would hinder innovation and disadvantage U.S. firms New York Raise Act N/A Preemption of state-level AI legislation In 2025, OpenAI supported federal preemption of state-level AI laws, arguing that a unified national framework would better promote innovation and avoid regulatory fragmentation.

Google DeepMind	**Update** **California SB 53** Industry group TechNet that represents Google opposed SB 53, arguing that the bill's scope is too broad and that the disclosure and reporting requirements could expose trade secrets or magnify security vulnerabilities. [Citizen Portal, 2025] [San Francisco Standard, 2025]	**Recap from Summer 2025** **EU AI Act** Google DeepMind opposed classifying general-purpose and foundational models as "high-risk," arguing this would stifle innovation and that regulation should target downstream applications. **US Legislations** California SB 1047 Google DeepMind opposed California's SB 1047, arguing that its safety rules would burden developers and stifle innovation and state oversight can fragment regulation. New York Raise Act Industry group with ties to Google opposed RAISE Act, arguing that the legislation could conflict with federal policy and impose overly broad restrictions on AI development. Preemption of state-level AI legislation In its response to the U.S. AI Action Plan in 2025, it called for federal leadership over issues like copyright, export controls, and development standards, warning that state-level rules could hinder innovation
Meta	**Update** **California SB 53** No public stance	**Recap from Summer 2025** **EU AI Act** Between 2022 and 2023, Meta lobbied EU institutions to limit safety rules in the AI Act, opposing strict obligations for general-purpose models and seeking exemptions for open-source systems. **US Legislations** California SB 1047 In 2024, Meta lobbied against California's SB 1047, arguing that its AI safety requirements—especially pre-deployment risk assessments and licensing—were overly broad and could hinder innovation New York Raise Act In 2025, Meta opposed RAISE Act through multiple affiliated groups, including Tech:NYC, the AI Alliance, and the Computer & Communications Industry Association. Preemption of state-level AI legislation In 2025, Meta advocated for federal preemption of state-level AI regulations, warning that fragmented laws could create compliance challenges and hinder innovation across jurisdictions
xAI	**Update** **California SB 53** No public stance	**Recap from Summer 2025** **EU AI Act** No public stance **US Legislations** California SB 1047 In 2024, xAI CEO Elon Musk publicly supported the bill in an X post. New York Raise Act No public stance Preemption of state-level AI legislation No public stance.
DeepSeek	DeepSeek is among the entities that have drafted the **Cybersecurity technology—Basic security requirements for generative artificial intelligence service**, which is a voluntary national standard that focuses on safety requirements including corpus safety, model safety, and safety assessment, although it doesn't mention frontier AI risks. [National Service Platform for Standards Information]	
Z.ai		
Alibaba Cloud	Alibaba is among the entities that have drafted the **Cybersecurity technology—Labeling method for content generated by artificial intelligence**, which is the only binding national standard that requires generative AI services to label AI-generated content both in implicit and explicit ways. [National Public Service Platform for Standards Information]	

TO BE COMPLETED BY PANELLISTS

Grading Sheet: Information Sharing and Public Messaging

Please pick a grade for each firm. You can add brief justifications to your grades.

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Grades								
Grade comments (Justifications, opportunities for improvements, etc.)								

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Provides detailed, verifiable disclosures on model safety and governance; fully cooperates with external evaluations; publicly and legislatively advocates for stronger safety and accountability standards.
- **B** Shares clear information on key safety and governance aspects; engages with external processes; publicly supports most safety initiatives while maintaining some self-interest.
- **C** Offers limited or curated safety and governance information; selectively participates in external efforts; adopts mixed or neutral positions on safety regulation.
- **D** Rarely discloses meaningful information; limited or inconsistent cooperation; messaging downplays risks or discourages stronger oversight.
- **F** Withholds or distorts safety information; no credible cooperation; messaging actively undermines safety regulation or misleads the public on risk.

Domain comments

Optional: Share observations that apply across companies, including general recommendations, notes on how you weighted indicators, or feedback on FLI's methodology.

Domain comments	

Appendix B: Company Survey

Introduction

Thank you for participating in the **FLI AI Safety Index 2025 Survey**. This survey is designed to allow your company to provide additional information about specific practices and policies for managing risks from advanced AI systems. The independent experts on the review panel will consider the information you provide here when evaluating your company's safety efforts.

Survey instructions

The survey contains a total of **34 questions**, which predominantly follow a **multiple-choice format**. Where options are provided, select the one that best fits your current practices. Some questions allow a brief explanation or ask for details (especially if you answered "Other" or an open-ended part) – please be concise and factual in those responses. You are welcome to provide **URLs or document references** for any publicly available policies or reports that support your answers. It is not necessary to answer all questions within the survey. You can skip specific questions when answering would be difficult/inconvenient.

You have received a personalized link which you can share with colleagues to collaborate on the survey. You do not need to fill out the survey in a single sitting. Progress will be saved whenever you navigate between sections.

Confidentiality

Please do not share confidential information. We plan to publish all survey responses in full after the grading process is completed.

We appreciate your time and effort in providing thorough answers.

Whistleblowing policies (16 Questions)

If your company has region-specific whistleblowing (WB) policies instead of a single global WB policy, please answer all questions in this survey with regard to the policy that applies to the majority of your frontier AI-focused management, research, and engineering employees. Unless a question specifically asks about other stakeholders, please answer based on protections available to current full-time employees. You may explain variations for different stakeholder groups in the final question.

You can use the text-box at the end of this section to provide clarifications and/or link to relevant publicly available documents.

Definition of terms:

Whistleblowing Function:

The organizational structure, personnel, processes, and resources established to receive, assess, investigate, and respond to whistleblowing reports. This includes the designated individuals or teams responsible for writing and acting according to the whistleblowing policy, managing the whistleblowing process, any technological systems used to facilitate reporting, and the mechanisms for investigating and addressing reported concerns.

Whistleblowing Policy:

The formal, documented set of rules, procedures, and guidelines that govern how an organization handles whistleblowing. This policy outlines what concerns can be reported ("material scope"), who can report them ("covered persons"), how reports should be made and to whom, how they will be handled, and what protections are available to whistleblowers who follow this policy. It serves as the official framework that defines the organization's approach to whistleblowing.

Covered persons:

Individuals who are explicitly protected when making good-faith reports under the whistleblowing policy. The range of covered persons may vary by organization and jurisdiction.

Material scope:

The range of issues, concerns, violations, or misconduct that can legitimately be reported through the whistleblowing channels and will be considered for investigation. In this context, this may include legal violations, ethical breaches, safety concerns, alignment issues, misrepresentations of capabilities, or other matters related to responsible AI development and deployment that the organization has defined as reportable concerns.

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Does your company have a WB policy & function covering frontier AI-focused staff? *Is this policy publicly accessible without login credentials?*	▪ Prefer not to answer (skips whistleblowing section) ▪ No WB policy & function - (skips whistleblowing section) ▪ Non-public policy exists - Please briefly explain your rationale for keeping it private:	Public WB policy - Please provide URL here: OpenAI's Raising Concerns Policy Blog Copy of Raising Concerns Policy (10.2024)	Non-public policy exists - Please briefly explain your rationale for keeping it private:	Prefer not to answer (skips whistleblowing section)	Non-public policy exists - Please briefly explain your rationale for keeping it private: Please see "post deployment monitoring" in our transparency hub. We expect to share more publicly in the near future. Anthropic's Transparency Hub: Voluntary Commitments	Public WB policy - Please provide URL here: Our code of conduct is public and we have several internal policies that cover whistleblowing. Google Code of Conduct
Who is formally designated with primary responsibility for overseeing the whistleblowing function and ensuring reports are properly addressed?	▪ Board/Audit Committee ▪ Executive management ▪ Compliance/Legal department ▪ HR department ▪ Other (Please also specify whom this role reports to):	▪ Board/Audit Committee ▪ Compliance/Legal department ▪ HR department	▪ Compliance/ Legal department			▪ Board/Audit committee ▪ Compliance/Legal and ▪ HR department

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Which statement best describes the investigative independence of your whistleblowing function?	▪ The whistleblowing function requires approval from management before initiating investigations based on whistleblower reports. ▪ The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management. ▪ The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.	The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.	The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.		The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.	The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.
Which of the following concerns are explicitly covered by your whistleblowing policy? (Select all that apply)	▪ Violations of applicable laws and regulations ▪ Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy) ▪ Credible safety concerns that may not violate specific policies including loss-of-control scenarios ▪ Pressure to compromise safety standards or suppress safety concerns ▪ Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices ▪ None of the above	▪ Violations of applicable laws and regulations ▪ Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy) ▪ Credible safety concerns that may not violate specific policies including loss-of-control scenarios ▪ Pressure to compromise safety standards or suppress safety concerns	▪ Violations of applicable laws and regulations		▪ Violations of applicable laws and regulations ▪ Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy) ▪ Credible safety concerns that may not violate specific policies including loss-of-control scenarios ▪ Pressure to compromise safety standards or suppress safety concerns ▪ Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices	▪ Violations of applicable laws and regulations ▪ Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy) ▪ Credible safety concerns that may not violate specific policies including loss-of-control scenarios ▪ Pressure to compromise safety standards or suppress safety concerns ▪ Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices
Does your whistleblowing policy explicitly protect individuals who report concerns in 'good faith' or with 'reasonable cause to believe', rather than requiring certainty that violations occurred?	▪ Yes ▪ No	Yes	Yes		Yes	Yes

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Which of the following persons are protected from retaliation under your whistleblowing policy? (Select all that apply)	▪ Current employees ▪ Former employees ▪ Contractors and self-employed workers ▪ AI research collaborators and academic partners ▪ Individuals who assist whistleblowers ▪ Suppliers and vendors with access to company systems	▪ Current employees ▪ Contractors and self-employed workers			▪ Current employees ▪ Former employees ▪ Contractors and self-employed workers ▪ AI research collaborators and academic partners ▪ Individuals who assist whistleblowers ▪ Suppliers and vendors with access to company systems	▪ Current employees ▪ Former employees ▪ Contractors and self-employed workers ▪ AI research collaborators and academic partners ▪ Individuals who assist whistleblowers ▪ Suppliers and vendors with access to company systems
To which of the following individuals or entities can whistleblowers submit reports according to your policy? (Select all that apply)	▪ Board member or board committee ▪ Dedicated Ethics/Whistleblowing Officer ▪ Ombudsperson ▪ Chief Compliance or Risk Officer ▪ General Counsel/Legal Department ▪ Human Resources department ▪ External/independent third party ▪ Direct disclosure to a statutory or supervisory authority ▪ Other (please briefly specify):	▪ Board member or board committee ▪ Chief Compliance or Risk Officer ▪ General Counsel/Legal Department ▪ Human Resources department ▪ External/independent third party ▪ Direct disclosure to a statutory or supervisory authority	▪ General Counsel/Legal Department ▪ Human Resources department ▪ Direct disclosure to a statutory or supervisory authority ▪ Other (please briefly specify): ▪ Manager			▪ Board member or board committee ▪ Dedicated Ethics/ Whistleblowing Officer ▪ Chief Compliance or Risk Officer ▪ General Counsel/Legal Department ▪ Human Resources department ▪ External/independent third party ▪ Direct disclosure to a statutory or supervisory authority
For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)	▪ Limited Reporting Channels ▪ Limited Reportable Issues ▪ Limited Retaliation Protection ▪ No Limitations For each, specify whether the limitation applies to: ▪ Former employees ▪ Contractors	▪ Limited Reporting Channels: Former employees ▪ Some channels, such as speaking to your current HR representative, are inherently available only to current employees.			▪ No Limitations: Former employees ▪ No Limitations: Contractors	▪ Limited Reporting Channels: Former employees ▪ Limited Reporting Channels: Contractors
Which of the following best describes the anonymity and confidentiality provisions in your whistleblowing policy? (Select the one that fits best)	▪ Our policy does not provide for anonymous reporting ▪ Our policy allows anonymous reporting but does not specify technical measures to protect reporter identity ▪ Our policy allows anonymous reporting with specific technical measures in place to protect reporter identity (e.g., anonymous hotline, encrypted system) ▪ Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports	Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports	Our policy allows anonymous reporting but does not specify technical measures to protect reporter identity		Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports	Our policy allows anonymous reporting with specific technical measures in place to protect reporter identity (e.g., anonymous hotline, encrypted system)

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Does your whistleblowing policy explicitly protect employees disclosing to external parties (e.g., regulators, accredited journalists, civil-society groups) when internal channels are unavailable, conflicted, or fail to resolve a serious concern within stated timelines? (Select one) *Possible Conditions:* ▪ *Imminent risk of serious harm* ▪ *Management or board implicated* ▪ *Reasonable fear of retaliation* ▪ *Internal investigation deadlines missed* ▪ *Unconditional reporting to a competent regulatory authority* ▪ *After internal reporting has been attempted*	▪ No – external disclosure is not explicitly protected or is discouraged (skips follow-up question) ▪ Limited – protected only under specific conditions (choose below) ▪ Full – broadly protected under all listed conditions above (skips follow-up question)	Full – broadly protected under all listed conditions above (skips follow-up question) Note: Our policy specifically protects disclosures to any "national, federal, state or local agency charged with the enforcement of any laws or regulations."	Limited – protected only under specific conditions (choose below)		Full – broadly protected under all listed conditions above (skips follow-up question)	Full – broadly protected under all listed conditions above (skips follow-up question)
If "Limited", under which circumstances is external disclosure protected?	▪ Imminent risk of serious harm ▪ Management or board implicated ▪ Reasonable fear of retaliation ▪ Internal investigation deadlines missed ▪ Unconditional reporting to a competent regulatory authority ▪ After internal reporting has been attempted ▪ Other (specify):		Unconditional reporting to a competent regulatory authority			
Which mechanisms ensure that your whistleblowing function has access to adequate (technical) expertise to investigate reports? (Select all that apply)	▪ Dedicated AI experts within the whistleblowing function itself ▪ Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary ▪ Standing agreements with external independent AI ethics/safety consultants ▪ Budget authority to engage external AI experts without requiring management approval ▪ None of the above ▪ Other (please specify):	▪ Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary				▪ Dedicated AI experts within the whistleblowing function itself ▪ Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary ▪ Budget authority to engage external AI experts without requiring management approval
Investigation timelines and escalation rights: Which best describes your policy's commitments? (Select one)	▪ None – no specific timelines for acknowledgment, updates, or resolution ▪ Basic – acknowledge receipt ≤ 7 days only ▪ Standard – acknowledge ≤ 7 days and provide updates ≤ 30 days ▪ Full – acknowledge ≤ 7 days, updates ≤ 30 days, final outcome ≤ 90 days ▪ Full + internal escalation – all Full timeframes plus whistleblowers may escalate to board/leadership if deadlines are missed ▪ Full + comprehensive escalation – all Full timeframes plus whistleblowers may escalate both internally AND to regulators/external parties if deadlines are missed	None – no specific timelines for acknowledgment, updates, or resolution	None – no specific timelines for acknowledgment, updates, or resolution			Full + comprehensive escalation – all Full timeframes plus whistleblowers may escalate both internally AND to regulators/external parties if deadlines are missed

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Which specific forms of retaliation are explicitly prohibited in your policy? (Check all that apply)	▪ Termination/Dismissal ▪ Demotion, or negative performance reviews ▪ Reduction in compensation or benefits ▪ Exclusion from meetings or information ▪ Harassment or creating a hostile work environment ▪ Blacklisting within the industry ▪ Legal action against the whistleblower ▪ None of the above	Our policy forbids retaliation. Notwithstanding the way this question is worded, it is well established under relevant law that retaliation can include termination or dismissal, demotion or negative performance reviews, or reduction in compensation or benefits. These are all covered under our policy's prohibition of retaliation. Our policy also expressly addresses harassment.	None of the above			▪ Termination/Dismissal ▪ Demotion, or negative performance reviews ▪ Reduction in compensation or benefits ▪ Exclusion from meetings or information ▪ Harassment or creating a hostile work environment ▪ Blacklisting within the industry ▪ Legal action against the whistleblower
Do any employment-, separation-, or settlement-related agreements used by your company contain non-disparagement or confidentiality clauses that could deter current or former employees from disclosing AI safety or risk-related concerns? (Select one)	▪ No - we do not include such restrictions in our agreements ▪ Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted. ▪ Yes, but not enforced – clauses exist, but the company has a written policy never to enforce (or threaten to enforce) them against AI safety or risk-related disclosures (no withholding of pay/equity and no legal action). ▪ Yes, enforced - our standard confidentiality and non-disparagement provisions may restrict raising AI safety or risk-related concerns	Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted. We have confidentiality clauses that could impact some forms of public disclosure, but these have carveouts for internal or regulator disclosures. We do not have non-disparagement clauses in any such agreements, except in specific cases where an employee or former employee has entered a mutual non-disparagement agreement with the company.			No - we do not include such restrictions in our agreements	Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted.
Which anti-retaliation provisions are explicitly detailed in your whistleblowing policy? (Select all that apply)	▪ Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation) ▪ Documented investigation procedure for retaliation claims (including designated investigators, timelines, evidence standards, and appeal rights) ▪ Concrete remedial measures for whistleblowers who experience retaliation (e.g., compensation, reinstatement, transfer options, or other specific remedies - not just general commitments to address retaliation) ▪ None of the above are specifically detailed	Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation)	None of the above are specifically detailed		Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation) Documented investigation procedure for retaliation claims (including designated investigators, timelines, evidence standards, and appeal rights)	Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation) Documented investigation procedure for retaliation claims (including designated investigators, timelines, evidence standards, and appeal rights)

External Pre-Deployment Safety Testing (6 Questions)

Please answer the following questions about external pre-deployment safety testing with regards to the release of your currently most capable publicly deployed AI model.

Frontier models:

- Anthropic - Claude Sonnet 4.5
- DeepSeek - R1
- Google Deepmind - Gemini 2.5 Pro
- Meta - Llama 4 Maverick
- OpenAI - GPT-5
- xAI - Grok-4
- Z.ai - GLM-4.6
- Alibaba Cloud - Qwen 3 Max

You can use the text-box at the bottom of the page to provide clarifications and/or link to relevant publicly available documents.

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Did your organisation commission one or more independent (no financial/ governance ties to your company) organisations to test this model for the dangerous capabilities or propensities you prioritized (in safety framework if available) before public release?	▪ No – no such external pre-deployment testing was commissioned (skip to next section) ▪ Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk"" (opens follow-up questions below):	Yes – external testing was commissioned. We've worked with the US CAISI and the UK AI Security Institute, independent third party labs such as METR, Apollo Research, SecureBio and Irregular Labs to add an additional layer of validation for key risks. Where possible and relevant, we report on their findings in our systems cards, such as in the GPT-5 System Card. Third party assessors were provided OpenAI GPT-5 Thinking early checkpoints, as well as the final launch candidate models to conduct their assessments across main preparedness categories (Cyber, Bio, AI Self-Improvement). As part of our ongoing efforts to consult with external experts, OpenAI granted early access to these versions of GPT-5 Thinking to both CAISI and UK AISI, both who conducted evaluations of the model's cyber and biological and chemical capabilities, as well as safeguards. As part of a longer-term collaboration, UK AISI was also provided access to prototype versions of our safeguards and information sources that are not publicly available – such as our monitor system design, biological content policy, and chains of thoughts of our monitor models. This allowed them to perform more rigorous stress testing and identify potential vulnerabilities more easily. Grey Swan and FAR.AI conducted general jailbreak red teaming. METR measured the model's general autonomous capabilities, and reward hacking, and Apollo Research evaluated in-context scheming and strategic deception. Pattern Labs evaluated the model's cybersecurity related capabilities, and SecureBio measured the models' biological capabilities.	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below):	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): CN CAICT: General Safety Issues	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): Please see our system cards (library, Claude Opus 4) and transparency hub for information on our external testing	Yes – external testing was commissioned. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., "UK AISI: cyber-offense, bio-risk (opens follow-up questions below): Yes, external safety testing was commissioned for 2.5, including across CBRN, Autonomy, Cyber, and Extremism and Radicalisation. We have worked with a diverse group of external experts, including Apollo Research, Dreadnode and Vaultis. (See link below) We plan to share more publicly about our approach. TIME - Exclusive: 60 U.K. Lawmakers Accuse Google of Breaking AI Safety Pledge

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
What was the highest level of technical access granted to any of the listed external evaluators during pre-deployment testing for the specified release? (Select the highest level that applies)	▪ Standard inference API with normal user-facing filters in place ▪ Inference API with safety filters disabled (no inference-time mitigations) ▪ Helpful-only" or base model API (no harmlessness fine-tuning and no filters) ▪ Fine-tuning interface without safety gatekeeping ▪ Direct read/write access to internal activations or weights	▪ Standard inference API with normal user-facing filters in place ▪ Inference API with safety filters disabled (no inference-time mitigations) ▪ Helpful-only" or base model API (no harmlessness fine-tuning and no filters)	Helpful-only" or base model API (no harmlessness fine-tuning and no filters)	"Helpful-only" or base model API (no harmlessness fine-tuning and no filters)		Inference API with safety filters disabled (no inference-time mitigations) External testing partners were provided the model without inference time mitigations relevant to their specific domain. We plan to set out more detail on our external testing programme in future.
What was the longest period of time that an external evaluator was given continuous access for pre-deployment testing of your model? (Select one)	▪ >5 weeks ▪ >3 weeks ▪ >2 weeks ▪ >1 week ▪ <1 week	>2 weeks	>5 weeks	>3 weeks		>3 weeks External testing partners began testing pre-deployment with interim findings provided before launch and then continued post deployment with further findings provided.
Which of the following publication arrangements applied to external evaluators' findings? *If different evaluators had different publication terms, please select all that occurred and briefly explain using the text-box.* *(select all that apply)*	▪ Evaluators may publish independently without prior company approval after the model is released. ▪ Evaluators may publish independently after company review/possible redaction. ▪ The company pre-committed to reproduce an independently written report in the model card without redactions. ▪ The company publishes report after review/possible redactions. ▪ The company provided its own summary of the evaluator's key findings. ▪ Findings remain internal ▪ Other: Please briefly explain:	Evaluators may publish independently without prior company approval after the model is released. ▪ This is true if they run their evaluations independently on the deployed model. Results from the pre-deployment evaluation period are under NDA / require prior approval to protect confidential information. Evaluators may publish independently after company review/possible redaction. ▪ See above, in cases where the evaluator wishes to publish about the specifics of the pre-deployment period - METR as an example did publish and made a note that they believe that our redactions did not substantively change their conclusions ("We did not make changes to conclusions, takeaways or tone (or any other changes we considered problematic) based on their review.") The company publishes report after review/possible redactions. ▪ OpenAI publishes excerpts from the report mutually agreed upon or written, with OpenAI having the final say for what content goes in System Cards. The company provided its own summary of the evaluator's key findings. ▪ This is true in some cases, but we also share back any summaries that we plan to publish with the evaluator prior to release to confirm factual accuracy.	Evaluators may publish independently after company review/possible redaction.	Evaluators may publish independently without prior company approval after the model is released.	Evaluators may publish independently after company review/possible redaction. The company provided its own summary of the evaluator's key findings.	The company provided its own summary of the evaluator's key findings. GDM publishes high level summaries appropriate for the risks being evaluated within the Models Cards / Tech report with GDM having the final say for what content goes in the Model Cards/ Tech report.

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
During pre-deployment testing, what best describes the query-rate or volume restrictions applied to external evaluators? *(Select one)*	▪ No limits – evaluators could automate or batch queries with no additional throttling or hard caps. ▪ Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). ▪ Public-tier caps – evaluators were held to the same rate/volume limits as ordinary paying users. ▪ Lower than Public-tier caps - evaluators had lower quotas than ordinary paying users.	Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). Query rates can depend on technical feasibility in some cases.	Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits).	No limits – evaluators could automate or batch queries with no additional throttling or hard caps.		Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits). Query rate is bespoke depending on the testing partner's specific needs and evaluation type. Where required, GDM provided elevated but capped quotas, but this rate often depended on technical feasibility.
Does your organization log and retain the model interactions of external evaluators during pre-deployment testing?	▪ Yes - Inputs and outputs are logged and retained. ▪ No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. ▪ Other (please describe):	Other (please describe): Zero Data Retention available upon request, if technically feasible during pre-deployment periods (for some new models or products, ZDR is not always possible during pre-deployment testing).	No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.	No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.		No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. No - Inputs and outputs are not logged during pre-deployment testing by external evaluators. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks.

Internal Deployments (3 Questions)

Deployment levels:

1. Broad deployment: Many teams within the company have access for normal use.
2. Development access: Access limited to specific teams or projects that are actively testing the model or developing it further.

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
If you specified external pre-deployment safety evaluations in the previous section, were these performed before or after broad internal deployment? (Select one)	▪ Before - External safety tests were completed before broad internal deployment. ▪ Partial - All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment. ▪ After - External safety tests were completed after broad internal deployment. ▪ Other (please explain briefly):	After - External safety tests were completed after broad internal deployment.	Before - External safety tests were completed before broad internal deployment.	Partial - All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment.		
What level of safety testing does your company require for broad internal deployment of frontier AI models? (Select one)	▪ No formal risk management requirements for internal deployments ▪ Formalized risk management for internal deployments with less stringent requirements than external deployment framework for the following risks/capabilities: situational awareness, scheming, AI R&D, cyber-offense. ▪ Formalized risk management for internal deployments with the same requirements as external deployment framework for the following risks/capabilities: situational awareness, scheming, cyber-offense. ▪ Company requires the same risk management effort for internal and external deployments. ▪ Other (Please briefly describe):	As described in our public Preparedness Framework, we believe that models that have reached or are forecasted to reach Critical capability under our framework will require additional safeguards (safety and security controls) during development, regardless of whether or when they are externally deployed. We do not currently possess any models that have Critical levels of capability, and we expect to further update this Preparedness Framework before reaching such a level with any model.	No formal risk management requirements for internal deployments	Company requires the same risk management effort for internal and external deployments.		
Does your company require any of the following safeguards for broad internal deployments of frontier AI models? *(Select all that apply)*	▪ Inference time safety mitigations for misuse risks (including cyber & bio risks) ▪ Restricting access to helpful-only models and only granting time-bound access to staff that apply with a legitimate research need ▪ Logging all inputs and outputs from internal use and retaining them for at least 30 days ▪ Not currently logging, but introduced an *official, written* plan to start doing so after models reach a specified capability threshold ▪ Analyzing all internal model interactions for abnormal activity, including harmful use or unexpected attempts by AI systems to take real-world actions ▪ Live monitoring and automated editing/ resampling of suspicious outputs ▪ None of the above ▪ Other (please describe briefly):	See answer to Q24, above.	Inference time safety mitigations for misuse risks (including cyber & bio risks) Restricting access to helpful-only models and only granting time-bound access to staff that apply with a legitimate research need	Restricting access to helpful-only models and only granting time-bound access to staff that apply with a legitimate research need Logging all inputs and outputs from internal use and retaining them for at least 30 days	We have nuanced rigorous approach to safeguards- each of these depends on product surface, classifier and harm type, and use case.	

Table continues on next page

Safety Practices, Frameworks, and Teams (9 Questions)

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
When you released your latest flagship model, did you release the same model version that the final round of safety (framework) evaluations were conducted on? (Select one)	▪ Yes – we released the same model version. ▪ No – we further modified the model but explicitly mentioned and described all further changes in the model documentation. ▪ No – further modifications are not described explicitly in the model documentation.	Yes – we released the same model version. Yes. All internal evaluations in the system card were conducted on the final checkpoint.	Yes – we released the same model version.	Yes – we released the same model version.	Yes – we released the same model version.	
If your company has one or more teams focused primarily on technical AI safety research, please provide more information about the team(s) below. By technical AI safety teams, we are referring to teams researching topics such as scalable oversight, dangerous capability evaluations, mechanistic interpretability, AI control, alignment evaluations, risk-modeling, etc. Please use separate paragraphs for listing multiple teams. 1) Team name (& website URL if available) 2) Mission and scope – Briefly describe the team's focus. Please distinguish between: ▪ immediate product safety (e.g., RLHF, jailbreak prevention, safety classifiers), and ▪ forward-looking/fundamental research (e.g., model organisms of misalignment, mechanistic interpretability) 3) Technical FTEs – Approximate number of full-time equivalent technical staff (researchers and research engineers). Please count each individual only once, based on their primary team.		We have multiple teams focused primarily on technical AI safety research, led by Johannes Heidecke (Safety Systems) and Mia Glaese (Alignment). Subteams and projects include: ▪ Mechanistic interpretability ▪ CoT interpretability ▪ Automating Alignment ▪ Safety oversight & control ▪ Dangerous capability evaluations ▪ Alignment evaluations ▪ Faithfulness & anti-scheming		1) Zhipu Evaluation Team & Zhipu Safety Team & Zhipu Posttraining Team We do not have team websites. 2) We prefer not to say. 3) 20~30	Aligned with our mission and origin as a safety research lab, we have multiple teams working on AI safety research including alignment science, interpretability, frontier red team, safeguards (research team, safeguards for Claude) and more.	
Does your organization have a formal, written policy that requires notifying external authorities when safety testing determines a model exceeds your organization's "unacceptable-risk" threshold (i.e., a risk-level that bars deployment under your own safety framework), even if the model will not be released? (Select option that best describes your policy)	1) No policy – there is no written requirement to notify any external body. 2) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority. 3) Regulator + public transparency – as in option 2 **and** the policy provides for a public statement or summary once doing so will not exacerbate security risks. 4) Other (please briefly describe):	1) No policy – there is no written requirement to notify any external body.	1) No policy – there is no written requirement to notify any external body.	1) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority.	Other (please briefly describe): U.S. Government notice when model requires ASL-3+ safeguards; see our RSP for more	
For companies that signed the ""Frontier AI Safety Commitments"" at the AI Seoul Summit in 2024, and those that strive to implement equivalent safety frameworks: Which of the levels below best describes the status of your Safety Framework? Please indicate the *highest* option below that accurately describes your current state.	▪ No official Safety Framework published (yet). ▪ Published & Implementation in progress ▪ Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet and the expected timeline for implementation: ▪ Published & fully implemented – All discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational.	Published & Implementation in progress We published version 2 of our Preparedness Framework on April 15 2025 and have implemented safeguards for high biological and chemical risk, which we first deployed with ChatGPT Agent, launched on July 17, 2025.	Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet and the expected timeline for implementation:	Published & Implementation in progress	Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet and the expected timeline for implementation:	

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Do you have a plan for ensuring that the AGI you're trying to build will remain controllable, safe and beneficial?	▪ No ▪ No, but we're working on it ▪ Yes, internally. (Please briefly explain why you have not published it)	Our mission is to ensure that artificial general intelligence benefits all of humanity. As part of our recently concluded recapitalization, the OpenAI Foundation became operational and has made an initial $25 billion commitment to invest in two areas: Health and curing disease, and technical solutions to AI resilience. For more on our approach to ensuring that AGI remains controllable and safe, see this post.	No	No, but we're working on it	Yes, publicly shared here (please provide URL): Anthropic, Responsible Scaling Policy, Version 2.2	
Which of the following elements of an AI emergency response capability has your organization implemented? (Select all that apply)	▪ Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months. ▪ Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. ▪ Conducted at least one full live emergency response drill/simulation in the past 12 months. ▪ Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander and a 24×7 duty roster. ▪ Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies. ▪ None of the above ▪ Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses	Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses OpenAI has developed and continues to improve incident response programs across key areas of its operations, including by improving and iterating on our AI safety incident-specific protocols that are tailored to our operations and technology. Our goal is to respond to incidents in a rapid, coordinated way. Our response capabilities include: ▪ Technical Controls for Rapid Mitigation: We maintain the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized within the last year in response to our finding that a GPT-4o model update was overly flattering or agreeable (see Sycophancy in GPT-4o: what happened and what we're doing about it, https://openai.com/index/sycophancy-in-gpt-4o/). ▪ Incident Response Planning and Structure: OpenAI has formal incident response plans for key areas of operations, including AI safety incident-specific protocols. Our response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. We maintain close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans. As part of our commitment to continuous improvement, we continue to refine our incident response capabilities, including robust playbooks for rapid-response. These efforts are integral to our broader model governance and safety assurance frameworks.		Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months. Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. Conducted at least one full live emergency response drill/simulation in the past 12 months. Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander and a 24 × 7 duty roster. Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.	Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses Please see our RSP and transparency hub for more	

Table continues on next page

Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
Does your company agree with the following principles for promoting legible and faithful reasoning in advanced AI systems to ensure AI remains safe and controllable? (Select all statements you support) Leading AI companies should:	▪ **Ensure Human-Legible Reasoning** - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. [No.] ▪ **Avoid Optimization That Encourages Obfuscation** - Developers should exercise caution when applying optimization pressures to model reasoning, especially ▪ when removing 'undesired reasoning', to prevent fostering deceptive behavior. ▪ **Disclose Optimization Pressures on Reasoning** - Companies should transparently report the optimization pressures and training methods applied to model reasoning, particularly when removing 'undesired reasoning.' ▪ None of the above	**Avoid Optimization That Encourages Obfuscation** - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior. We've published research and joined a broader working paper urging against optimizing on chains of thought: As we noted in the GPT-5 system card, "our commitment to keep our reasoning models' CoTs as monitorable as possible (i.e., as faithful and legible as possible) allows us to conduct studies into our reasoning models' behavior by monitoring their CoTs."		**Ensure Human-Legible Reasoning** - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. **Avoid Optimization That Encourages Obfuscation** - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior.		
Task-Specific Fine-Tuning (TSFT) involves training a model to excel at potentially dangerous tasks (e.g., designing biological agents, cyber attacks). Before releasing your current frontier model, which statement best describes your TSFT safety testing? (Select one)	▪ None – no TSFT safety testing performed (skips follow-up). ▪ Partial – TSFT performed on ≤ 2 high-risk domains (choose below). ▪ Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below).	None for gpt-5. We evaluated helpful-only models, which we believe is appropriate for the threat model of misuse for models made available via our platform and whose weights we do not release, as is codified in our Preparedness Framework. Note that we did task-specific fine tuning on biological and cyber capabilities for gpt-oss and published a paper with our findings, Estimating worst case frontier risks of open weight LLMs.		Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below).		
If you selected 'Partial' or 'Comprehensive' on the previous question, Please tick the risk-domains tested with TSFT.	▪ Biological ▪ Persuasion ▪ Chemical ▪ Deceptive alignment / Autonomy ▪ Cyber-offense ▪ Other (please specify):			Deceptive alignment / Autonomy		
If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number. You may also share additional information about your company's policies.		Below, we include some additional information about our security work that we believe may be useful context for evaluators considering our overall posture and approach. ▪ For additional technical detail on our security measures for AI see: Security on the path to AGI ▪ Third party collaboration on security: OpenAI maintains a bug bounty program through BugCrowd, and welcomes responsible disclosures from third parties via our coordinated vulnerability disclosure policy. In addition, OpenAI runs a Cybersecurity Grant Program to support research and development focused on protecting AI systems and infrastructure. This program encourages and funds initiatives that help identify and address vulnerabilities, ensuring the safe deployment of AI technologies.				

FLI AI Safety Index

Independent experts evaluate safety practices of leading AI companies across critical domains.

December 2025