Future of Life Institute

AI Safety Index

Summer 2025

17th July 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 seven 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 second 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	x.AI	Meta	Zhipu AI	DeepSeek
Overall Grade	C+	C	C-	D	D	F	F
Overall Score	2.64	2.10	1.76	1.23	1.06	0.62	0.37
Risk Assessment	C+	C	C-	F	D	F	F
Current Harms	B-	B	C+	D+	D+	D	D
Safety Frameworks	C	C	D+	D+	D+	F	F
Existential Safety	D	F	D-	F	F	F	F
Governance & Accountability	A-	C-	D	C-	D-	D+	D+
Information Sharing	A-	B	B	C+	D	D	F

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

- **Anthropic gets the best overall grade (C+).** The firm led on risk assessments, conducting the only human participant bio-risk trials, excelled in privacy by not training on user data, conducted world-leading alignment research, delivered strong safety benchmark performance, and demonstrated governance commitment through its Public Benefit Corporation structure and proactive risk communication.
- **OpenAI secured second place ahead of Google DeepMind.** OpenAI distinguished itself as the only company to publish its whistleblowing policy, outlined a more robust risk management approach in its safety framework, and assessed risks on pre-mitigation models. The company also shared more details on external model evaluations, provided a detailed model specification, regularly disclosed instances of malicious misuse, and engaged comprehensively with the AI Safety Index survey.
- **The industry is fundamentally unprepared for its own stated goals.** Companies claim they will achieve artificial general intelligence (AGI) within the decade, yet none scored above D in Existential Safety planning. One reviewer called this disconnect "deeply disturbing," noting that despite racing toward human-level AI, "none of the companies has anything like a coherent, actionable plan" for ensuring such systems remain safe and controllable.
- **Only 3 of 7 firms report substantive testing for dangerous capabilities linked to large-scale risks such as bio- or cyber-terrorism** (Anthropic, OpenAI, and Google DeepMind). While these leaders marginally improved the quality of their model cards, one reviewer warns that the underlying safety tests still miss basic risk-assessment standards: "The methodology/reasoning explicitly linking a given evaluation or experimental procedure to the risk, with limitations and qualifications, is usually absent. [...] I have very

low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further."

- **Capabilities are accelerating faster than risk management practice**, and the gap between firms is widening. With no common regulatory floor, a few motivated companies adopt stronger controls while others neglect basic safeguards, highlighting the inadequacy of voluntary pledges.
- **Whistleblowing policy transparency remains a weak spot.** Public whistleblowing policies are a common best practice in safety-critical industries because they enable external scrutiny. Yet, among the assessed companies, only OpenAI has published its full policy, and it did so only after media reports revealed the policy's highly restrictive non-disparagement clauses.
- **Chinese AI firms Zhipu.AI and DeepSeek both received failing overall grades.** However, the report scores companies on norms such as self-governance and information-sharing, which are far less prominent in Chinese corporate culture. Furthermore, as China already has regulations for advanced AI development, there is less reliance on AI safety self-governance. This is in contrast to the United States and United Kingdom, where the other companies are based, and which have, as yet, passed no such regulation on frontier AI.
    - *Note: the scoring was completed in early July and does not reflect recent events such as xAI's Grok 4 release, Meta's superintelligence announcement, or OpenAI's commitment to sign the EU AI Act Code of Practice.*

These findings reveal an unregulated industry where competitive pressures and technological ambition far outpace safety infrastructure and norms. This imbalance becomes more dangerous as companies pursue their stated goal of achieving artificial general intelligence (AGI) that matches or exceeds human capabilities within the decade.

1.2 Improvement opportunities by company

Here we highlight examples of how individual companies can improve future scores with relatively modest effort.

Anthropic:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Become more transparent and explicit about risk assessment methodology–e.g. why/how exactly is the particular eval related to a (class of) risks. Include reasoning in model cards that explicitly links evaluations or experimental procedures to specific risk, with limitations and qualifications.

OpenAI:

- Rebuild lost safety team capacity and demonstrate renewed commitment to OpenAI's original mission.
- Maintain the strength of current non-profit governance elements to guard against financial pressures undermining OpenAI's mission.

Google DeepMind:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Publish evaluation results for models without safety guardrails to more closely approximate true model capabilities.
- Improve coordination between DeepMind safety team and Google's policy team.
- Increase transparency around and investment in third-party model evaluations for dangerous capabilities.

xAI:

- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Boost current draft safety framework to match the efforts by Anthropic and OpenAI.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Meta:

- Significantly increase investment in technical safety research, especially tamper-resistant safeguards for open-weight models.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Zhipu AI:

- Publish the AI Safety Framework promised at the AI Summit in Seoul.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.

DeepSeek:

- Address extreme jailbreak vulnerability before next release.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Develop and publish a comprehensive AI safety framework.

All companies: Publish a first concrete plan, however imperfect, for how they hope to control the AGI/ASI they plan to build.

1.3 Methodology

Index Structure: The 2025 Index evaluates seven leading AI companies on 33 indicators spanning six critical domains.

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 Model Vulnerabilities

Current Harms

Model Safety / Trustworthiness
- Stanford's HELM Safety Benchmark
- Stanford's HELM AIR Benchmark
- TrustLLM Benchmark

Robustness
- Gray Swan Arena: UK AISI Agent Red-Teaming Challenge
- Cisco Security Risk Evaluation
- Protecting Safeguards from Fine-tuning

Digital Responsibility
- 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

Lobbying on AI Safety Regulations

Company Structure & Mandate

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

Information Sharing

Technical Specifications
- System Prompt Transparency
- Behavior Specification Transparency

Voluntary Cooperation
- G7 Hiroshima AI Process Reporting
- FLI AI Safety Index Survey Engagement

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

Data Collection: Evidence was gathered between March 24 and June 24, 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. The complete evidence base is documented in Appendix A and Appendix B.

Expert Evaluation: An independent panel of leading AI researchers and governance experts reviewed company-specific evidence and assigned domain-level grades (A-F) based on absolute performance standards. 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:

Dylan Hadfield-Menell

Dylan Hadfield-Menell is the Bonnie and Marty (1964) Tenenbaum Career Development 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.

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.

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 also serves as the Director of the UC Berkeley AI Policy Hub, an expert in the OECD Expert Group on AI Risk and Accountability, and a member of the U.S. AI Safety Institute Consortium.

Sneha Revanur

Sneha Revanur is the founder and president of Encode, a global youth-led 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.

Stuart Russell

Stuart Russell is a Professor of Computer Science at the University of California at Berkeley, holder of the Smith-Zadeh Chair in Engineering, 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 co-authored the standard textbook for AI, which is used in over 1500 universities in 135 countries.

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 artificial intelligence through technical research as well as education, outreach, governance and advocacy.

2 Introduction

Artificial intelligence systems are becoming more capable and autonomous at an unprecedented pace. Fueled by massive investments and technical breakthroughs, these general-purpose AI systems are transforming from specialized tools into increasingly versatile agents, being deployed in increasingly high-stakes settings. These trends pose significant risks, ranging from malicious use to systemic failures and loss of meaningful human control. This makes independent scrutiny of how GPAI systems are developed and deployed more urgent than ever.

The AI Safety Index is a response to that urgency. Developed and published by the Future of Life Institute (FLI), in collaboration with a review panel composed of some of the foremost independent experts in AI, the Index provides an independent assessment of how responsibly the world's leading AI companies are developing and deploying increasingly powerful AI systems. It focuses on institutional safeguards, such as risk management frameworks, third-party oversight, and whistleblower policies. Each company is evaluated on a set of 33 indicators across six domains, and scores are presented in a format designed to be accessible to both expert and non-expert audiences.

Since the release of the inaugural Index in December 2024, the global awareness around AI risks has continued to increase. Mandated by the nations attending the AI Safety Summit in the UK, the first International AI Safety Report synthesized the current scientific understanding of AI risks and potential mitigation strategies, acknowledging the potential for catastrophic outcomes. Distinguished subject experts from industry, government, and academia convened in Singapore and agreed on the Singapore Consensus on Global AI Safety Research Priorities, which outlines key research priorities for addressing risks from advanced AI. These developments confirm a growing international consensus: keeping AI safe as capabilities advance demands urgent investment in alignment research and substantial improvements in risk management.

The Summer 2025 edition of the FLI AI Safety Index evaluates seven frontier AI companies—OpenAI, Anthropic, Google DeepMind, Meta, xAI, Zhipu AI, and, for the first time, DeepSeek—using updated and improved indicators that reflect evolving deployment practices and safety norms. The Index is intended to be a practical, public-facing tool for tracking corporate behavior, identifying emerging best practices, and surfacing critical gaps. By making companies' risk management practices more visible and comparable, the Index aims to strengthen incentives for responsible AI development and to close the gap between safety commitments and real-world actions.

3 Methodology

Evaluating safety practices at the cutting edge of AI development requires a flexible and evolving methodology that can capture and appropriately weigh both concrete implementations and stated positions, commitments, and levels of transparency across diverse organizations.

This section outlines how the AI Safety Index evaluates and grades AI companies. We explain indicator design and structure, our evidence collection and data sources, the independent review process, and discuss notable limitations. By making our methods fully transparent, we aim to provide stakeholders with the context required to understand both the strengths and limitations of our assessments.

3.1 Companies Assessed

The 2025 FLI AI Safety Index evaluates seven leading general-purpose AI developers: Anthropic, OpenAI, Google DeepMind, Meta, Zhipu AI, x.AI, and DeepSeek. These companies represent the global frontier of AI capability development and were selected based on flagship model performance. The inclusion of the Chinese firms Zhipu AI and DeepSeek also reflects our intention to make the Index representative of leading developers globally. The inclusion of DeepSeek for this second iteration recognizes both its technical achievements in efficient model training and its growing influence in the Chinese AI ecosystem. Future iterations of the Index may assess different companies as the competitive landscape evolves.

Our selection focused on firms that have deployed models with competitive performance on public benchmarks. Therefore, we excluded companies that aim to build artificial general intelligence but have not yet deployed frontier models, such as Safe Superintelligence Inc.

3.2 Index Design and Structure

The AI Safety Index evaluates companies on a set of 33 indicators across six domains that capture different aspects of responsible AI development and deployment. Each domain contains multiple indicators along which differences in responsible AI practices between firms become visible. The indicators were selected based on the five criteria below.

- **High signal value:** Revealing substantive differences in companies' investments and priorities
- **Implementation focus:** Prioritizing demonstrated measures overstated commitments
- **Information availability:** Ensuring sufficient public evidence exists for evaluation
- **Clear definition:** Enabling consistent evaluation across companies
- **Leadership recognition:** Rewarding exceptional practices while maintaining adequate standards

Complete indicator definitions, rationales, and sources are provided in Appendix A. Below, we briefly introduce each of the 33 indicators:

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 Model 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.

Model Safety / Trustworthiness	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.
Robustness	Gray Swan Arena (UK AISI Agent Red-Teaming Challenge)	Tests how AI agents withstand adversarial challenges in high-risk, simulated decision-making environments to expose vulnerabilities.
	Cisco Security Risk Evaluation	Reports cybersecurity assessments of AI systems, focusing on their vulnerability to automated jailbreaking attacks.
	Protecting Safeguards from Fine-tuning	Evaluates whether AI providers implement protections that prevent fine-tuning from disabling important safety mechanisms or filters.
Digital Responsibility	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 and 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 audits whether each company's governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems. Indicators examine whistleblowing systems, legal structures, and advocacy on AI regulations.

Lobbying on AI Safety Regulations		Tracks how a company has engaged in lobbying or public advocacy that influences laws and policies related to AI safety.
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.
	FLI AI Safety Index Survey Engagement	Reports that companies voluntarily completed and submitted FLI's detailed safety survey to supplement publicly available information.
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.

3.3 Related Work and Incorporated Research

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, and two projects by Zach Stein-Perlman—AILabWatch.org and AISafetyClaims.org—which provide detailed, technical evaluations of how leading AI companies work to avert catastrophic risks from advanced AI.

Incorporated Research: Where appropriate, the 2025 Index incorporates existing comparative analysis led by credible research institutions. The Safety Framework domain imports SaferAI's in-depth assessment of firms' published safety frameworks in the 'Safety Framework' domain. SaferAI is a leading governance and research non-profit with significant expertise in AI risk management. The Index further incorporates AILabWatch.org's tracker of technical AI safety research in the 'Existential Safety' domain. 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, and the HELM AIR-Bench and HELM Safety benchmarks by Stanford's Center for Research on Foundation Models. The section further features results from the UK AI Security Institute's Red-teaming Challenge on the Gray Swan Arena, and a model security analysis from Cisco.

3.4 Data Sources and Evidence Collection

The evidence collection process for the 2025 AI Safety Index was conducted between March 24 and June 24, 2025, using 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.

Desk research: Our primary evidence base 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 May 28, and firms were given until June 17 to respond. The survey can be reviewed in full in Appendix B. Compared to the previous winter 2024 iteration of the Index, the updated survey was shorter and more specifically targeted 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 three companies (OpenAI, Zhipu AI, and xAI), representing 43% of assessed firms. Anthropic, Google DeepMind, Meta, and DeepSeek have not submitted a response. Full survey responses are attached in Appendix B as well.

Grading Sheets: The resulting evidence base underlying the index was then structured into the grading sheets found in Appendix A. The grading sheets, which are split into six domains, contain company-specific information for each of the 33 indicators of the current edition of the index. For each indicator, the grading

sheets present a definition of its scope, a rationale for the inclusion of the indicator, and references to relevant literature where appropriate. The sources for all pieces of company-specific information are embedded in the relevant locations with hyperlinks. We prioritized primary sources directly from companies over secondary reporting wherever possible, with investigative journalism providing valuable insights, uncovering practices not voluntarily disclosed. Where indicators overlap with survey questions, relevant survey responses were highlighted in the grading sheets.

3.5 Grading Process and Expert Review

The 2025 AI Safety Index's grading process was designed to ensure an impartial and qualified evaluation of the companies' performance across the selected indicators. It features a review panel of distinguished independent experts who assess the company-specific evidence for each indicator 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 six 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. The panel's composition further reflects a diversity of backgrounds, given that the Index spans from technical AI Safety topics to the domains of governance and policy. The panel thus features both renowned machine learning professors who specialize in alignment and control, and also governance experts from the non-profit sector. The composition of the panel remained mostly unchanged from the previous version of the index. We are grateful to Dylan Hadfield-Menell for joining the panel as a new member, replacing Yoshua Bengio, who stepped back due to time constraints from competing professional commitments. Review panel member Atoosa Kasirzadeh had to pause her contribution for the current version due to a family emergency. 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 June 24, and the grading period ended on July 9. After reviewing the evidence, reviewers assigned letter grades (A+ to F) to each company per domain. For each grade, 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 a given domain. Overall grades represent the average across all domains.

3.6 Limitations

Our methodology has several important limitations that should be considered when interpreting the Index results.

Information Availability and Verification

Our evaluation relies primarily on public information, which creates fundamental constraints. Companies control their disclosure levels, making it difficult to distinguish between poor transparency and poor 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.

The 33 indicators represent a subset of important practices for which meaningful evidence exists, not a comprehensive assessment of all safety dimensions. Furthermore, we cannot independently verify company claims and must assume official reports are truthful—a significant limitation given the high stakes involved.

Geographic and Cultural Context

Our methodology, developed in Western academic contexts, is rather Western-centric and may adversely impact the scores of the Chinese companies Zhipu and DeepSeek. For example, it places a premium on self-governance and information-sharing, both of which are far less prominent in Chinese corporate culture. As China already has regulations for generative AI in place, firms face less incentive to self-regulate when it comes to AI safety. This is in contrast to the US and UK, where the other companies are based, and where no such regulation exists

Moreover, information availability varies dramatically across regions. U.S.-based companies sometimes provide extensive documentation, from detailed model cards to public safety frameworks. Chinese companies such as Zhipu AI and DeepSeek operate under different regulatory frameworks and cultural norms around transparency, making direct comparisons challenging. Language barriers compound these challenges, potentially affecting our assessment of non-English resources. Several indicators have limited applicability to Chinese firms operating in fundamentally different contexts.

Methodological Constraints

Our focus on observable, documentable practices may undervalue crucial but hard-to-measure factors such as safety culture. Additionally, while our six-member panel brings diverse expertise, it cannot encompass all relevant domains. Reviewers' backgrounds inevitably influence assessments, and the flexibility in weighting indicators may introduce inconsistencies.

Moving Forward

We aim to mitigate these limitations through rigorous documentation of sources, methodology, and reviewer materials. Readers should interpret Index results as one input among many for understanding AI safety practices. We invite constructive criticism and helpful suggestions at policy@futureoflife.org and are committed to improving the project with every iteration.

4 Results

Overall Rankings: Anthropic leads with a C+ (2.64), followed by OpenAI (C, 2.10) and Google DeepMind (C-, 1.76). The middle tier includes x.AI and Meta (both D), while Chinese companies Zhipu AI and DeepSeek trail with failing grades. Notably, no company achieved higher than C+, indicating that even industry leaders fall short of adequate safety standards.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Overall Grade	C+	C	C-	D	D	F	F
Overall Score	2.64	2.10	1.76	1.23	1.06	0.62	0.37
Risk Assessment	C+	C	C-	F	D	F	F
Current Harms	B-	B	C+	D+	D+	D	D
Safety Frameworks	C	C	D+	D+	D+	F	F
Existential Safety	D	F	D-	F	F	F	F
Governance & Accountability	A-	C-	D	C-	D-	D+	D+
Information Sharing	A-	B	B	C+	D	D	F

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

- **Anthropic gets the best overall grade (C+).** The firm led on risk assessments, conducting the only human participant bio-risk trials, excelled in privacy by not training on user data, conducted world-leading alignment research, delivered strong safety benchmark performance, and demonstrated governance commitment through its Public Benefit Corporation structure and proactive risk communication.
- **OpenAI secured second place ahead of Google DeepMind.** OpenAI distinguished itself as the only company to publish its whistleblowing policy, outlined a more robust risk management approach in its safety framework, and assessed risks on pre-mitigation models. The company also shared more details on external model evaluations, provided a detailed model specification, regularly disclosed instances of malicious misuse, and engaged comprehensively with the AI Safety Index survey.
- **The industry is fundamentally unprepared for its own stated goals.** Companies claim they will achieve artificial general intelligence (AGI) within the decade, yet none scored above D in Existential Safety planning. One reviewer called this disconnect "deeply disturbing," noting that despite racing toward human-level AI, "none of the companies has anything like a coherent, actionable plan" for ensuring such systems remain safe and controllable.
- **Only 3 of 7 firms report substantive testing for dangerous capabilities linked to large-scale risks such as bio- or cyber-terrorism** (Anthropic, OpenAI, and Google DeepMind). While these leaders marginally improved the quality of their model cards, one reviewer warns that the underlying safety tests still miss basic risk-assessment standards: "The methodology/reasoning explicitly linking a given evaluation or experimental procedure to the risk, with limitations and qualifications, is usually absent. [...] I have very low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further."

- **Capabilities are accelerating faster than risk management practice**, and the gap between firms is widening. With no common regulatory floor, a few motivated companies adopt stronger controls while others neglect basic safeguards, highlighting the inadequacy of voluntary pledges.
- **Whistleblowing policy transparency remains a weak spot.** Public whistleblowing policies are a common best practice in safety-critical industries because they enable external scrutiny. Yet, among the assessed companies, only OpenAI has published its full policy, and it did so only after media reports revealed the policy's highly restrictive non-disparagement clauses.
- **Chinese AI firms Zhipu.AI and DeepSeek both received failing overall grades.** However, the report scores companies on norms such as self-governance and information-sharing, which are far less prominent in Chinese corporate culture. Furthermore, as China already has regulations for advanced AI development, there is less reliance on AI safety self-governance. This is in contrast to the United States and United Kingdom, where the other companies are based, and which have, as yet, passed no such regulation on frontier AI.
    - *Note: the scoring was completed in early July and does not reflect recent events such as xAI's Grok 4 release, Meta's superintelligence announcement, or OpenAI's commitment to sign the EU AI Act Code of Practice.*

These findings reveal an unregulated industry where competitive pressures and technological ambition far outpace safety infrastructure and norms. This imbalance becomes more dangerous as companies pursue their stated goal of achieving artificial general intelligence (AGI) that matches or exceeds human capabilities within the decade.

4.2 Improvement opportunities by company

Here we highlight examples of how individual companies can improve future scores with relatively modest effort.

Anthropic:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Become more transparent and explicit about risk assessment methodology–e.g. why/how exactly is the particular eval related to a (class of) risks. Include reasoning in model cards that explicitly links evaluations or experimental procedures to specific risk, with limitations and qualifications.

OpenAI:

- Rebuild lost safety team capacity and demonstrate renewed commitment to OpenAI's original mission.
- Maintain the strength of current non-profit governance elements to guard against financial pressures undermining OpenAI's mission.

Google DeepMind:

- Publish a full whistleblowing policy to match OpenAI's transparency standard.
- Publish evaluation results for models without safety guardrails to more closely approximate true model capabilities.
- Improve coordination between DeepMind safety team and Google's policy team.
- Increase transparency around and investment in third-party model evaluations for dangerous capabilities.

xAI:

- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Boost current draft safety framework to match the efforts by Anthropic and OpenAI.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Meta:

- Significantly increase investment in technical safety research, especially tamper-resistant safeguards for open-weight models.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Publish a full whistleblowing policy to match OpenAI's transparency standard.

Zhipu AI:

- Publish the AI Safety Framework promised at the AI Summit in Seoul.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.

DeepSeek:

- Address extreme jailbreak vulnerability before next release.
- Ramp up risk assessment efforts and publish implemented evaluations in upcoming model cards.
- Develop and publish a comprehensive AI safety framework.

All companies: Publish a first concrete plan, however imperfect, for how they hope to control the AGI/ASI they plan to build.

4.3 Domain-level findings

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.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	C+	C	C-	F	D	F	F
Score	2.5	2.2	1.8	0	1.0	0.35	0

Indicator overview

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 Model Vulnerabilities

Only three companies–Anthropic, Google DeepMind, and OpenAI–were found to show meaningful efforts to assess whether their models pose large-scale risks. Reviewers recognized these efforts as demonstrating a substantial investment, highlighting Anthropic's and OpenAI's assessment of helpful-only models without safety guardrails as notable best practice. The review panel furthermore commended Anthropic as the only company to conduct a human participant uplift trial to evaluate the impact of its flagship model on risks from bioterrorism. The review panel found that the remaining four companies lack basic risk assessment documentation for critical risks, with basic practices such as dangerous capability evaluations and model cards mostly absent.

However, even the leaders were **not deemed to be sufficiently rigorous** in their approaches by the review panel. One expert criticized the lack of methodological transparency, noting: "The methodology/reasoning explicitly linking a given evaluation or experimental procedure to the risk, with limitations and qualifications, is usually absent." The reviewer encouraged the companies to draw from the risk assessment literature and improve their approach by being more "transparent and explicit about their risk assessment methodology (e.g., why/how exactly is the particular eval related to a (class of) risks". Currently, reported assessments were found to feature little explanation of why specific evaluations were chosen, what risks they target, or how results should be interpreted.

Reviewers also expressed concerns that none of the companies commissioned independent verifications or assessments of internal safety evaluations, which means reported evidence needs to be accepted on trust. While OpenAI and Anthropic tasked competent external evaluators to assess some risks, reviewers noted that these efforts provide limited assurance, as evaluators are made to sign NDAs. Highlighting the severity of industry-wide deficiencies, one panellist who, despite assigning Anthropic the highest score among all firms, concluded her assessment by stating: "I have very low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evals decreases my confidence further."

Current Harms

Companies were evaluated on how effectively their models mitigate current harms, with a focus on safety benchmark performance, robustness against adversarial attacks, watermarking of AI-generated content, and the treatment of user data.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	B-	B	C+	D+	D+	D	D-
Score	2.8	3.0	2.5	1.5	1.65	1.0	0.85

Indicator overview

Model Safety / Trustworthiness
- Stanford's HELM Safety Benchmark
- Stanford's HELM AIR Benchmark
- TrustLLM Benchmark

Robustness
- Gray Swan Arena: UK AISI Agent Red-Teaming Challenge
- Cisco Security Risk Evaluation
- Protecting Safeguards from Fine-tuning

Digital Responsibility
- Watermarking
- User Privacy

The review panel found that performance on safety benchmarks varies drastically, from B's to D's. All models remain vulnerable to jailbreaks and misuse, with DeepSeek standing out for particularly high failure rates in adversarial testing. Reviewers positively noted the incremental improvements in model robustness showcased by Anthropic and OpenAI. The panel criticized Meta, ZhipuAI, and DeepSeek for further amplifying risks by fully releasing their model weights, which enables malicious actors to remove safety protections through fine-tuning.

Watermarking systems were found to remain underdeveloped for most companies, despite their importance in addressing the harms of synthetic content. Google DeepMind's SynthID stood out as the most advanced implementation.

On privacy matters, Anthropic was highlighted as the only firm not to train on user interaction data by default. Reviewers acknowledged that Meta, Zhipu, and DeepSeek offer users the ability to self-host their AIs by sharing model weights, which enables the highest level of privacy.

Safety Frameworks

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

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	C	C	D+	D+	D+	F	F
Score	2.15	2.0	1.5	1.5	1.5	0	0

Indicator overview

Risk Identification | **Risk Analysis and Evaluation** | **Risk Treatment** | **Risk Governance**

ZhipuAI and DeepSeek were assigned Fs for not having published a comparable Safety Framework in the first place, even though ZhipuAI had promised to do so at the international AI summit in Seoul. The remaining firms

published frameworks, with reviewers highlighting distinct strengths: Anthropic's risk identification and mitigation approach, OpenAI's commitment to publishing evaluation results, Google DeepMind's alert thresholds, and Meta's provisions for ongoing monitoring and threat modeling. However, reviewers identified one overarching criticism: the very limited scope of risks addressed by these frameworks.

The absence of external oversight mechanisms has emerged as a fundamental weakness of current frameworks, with OpenAI and Anthropic being noted for their early efforts to include external stakeholders. One reviewer explained that they emphasized risk treatment and governance in their weighting because: "[..] if no one has point and power [...], the quality of risk understanding is kind of moot." The panel further pointed out that no framework sufficiently defined specifics around conditional pauses. While firms signed the Seoul commitments, none have spelled out concrete, externally verifiable trigger thresholds for pauses, nor reliable enforcement mechanisms.

Overall, panellists concluded that none of the companies could be trusted to prevent catastrophic risks with a high degree of confidence. Experts found that while the quality of these voluntary frameworks is slowly improving, they lack critical governance mechanisms that can ensure frameworks are implemented and enforced in high-stakes situations.

SaferAI's complete evaluation of firms' safety frameworks contains additional analysis from risk management professionals and can be found here. Differences in scoring are due to the weightings applied by our review panel.

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.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	D	F	D-	F	F	F	F
Score	1.0	0.67	0.77	0.23	0.33	0	0

Indicator overview

Existential Safety Strategy	Internal Monitoring and Control Interventions	Technical AI Safety Research	Supporting External Safety Research

"This category is deeply disturbing," one reviewer noted. All seven companies are racing to build AGI within the decade, yet "literally none of the companies has anything like a coherent, actionable plan for what should happen if what they say will happen soon and are very actively working to make happen, happens". Multiple reviewers emphasized this stark disconnect between ambition and preparedness, with five of seven companies scoring an F, and none of them scoring better than a D.

Quantitative guarantees for alignment or control strategies were found to be virtually absent, with no firm providing formal safety proofs or probabilistic risk bounds for the transformative technologies they set out to develop. "Companies working on AGI need to show that risks are actually below an acceptable threshold. None of them have a plan to do this," one reviewer highlighted, adding that even those showing awareness are pursuing approaches "unlikely to yield the necessary level of safety."

Anthropic and Google DeepMind received a D and a D-. One reviewer highlighted Anthropic's "world-leading research on scheming / alignment faking, which lends credibility to their commitment to detecting misalignment." However, the panel criticised the firm's strategy's over-reliance on mechanistic interpretability, given that the discipline is in an early stage.

Google DeepMind's safety documentation was described as "well thought out, with a serious commitment to monitoring," but reviewers noted it provides no solid foundation for assessing risks as acceptably low.

OpenAI's deteriorating safety culture drew particular concern, leading to a grade drop to an F. "OpenAI's focus on safety has decreased over the last year, and it has lost most of its researchers in this area," a reviewer noted. High turnover on the safety team and failure to meet Superalignment commitments were taken as an indication of a concerning shift in priorities.

While xAI's and ZhipuAI's leadership was acknowledged for showing awareness of catastrophic risks, the companies themselves produce little concrete technical research aimed at addressing these risks. DeepSeek was also found not to publish related technical research.

Governance and Accountability

This domain assesses whether each company's governance structure and day-to-day operations prioritize meaningful accountability for the real-world impacts of its AI systems. Indicators examine whistleblowing systems, legal structures, and advocacy efforts related to AI regulations.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	A-	C-	D	C-	D-	D+	D+
Score	3.7	1.7	1.0	1.85	0.85	1.35	1.35

Indicator overview

Lobbying on AI Safety Regulations

Company Structure & Mandate

Whistleblowing

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

Anthropic stood out to reviewers for its Public Benefit Corporation status and Long-Term Benefit Trust, designed to reduce short-term profit incentives and strengthen long-term safety considerations. The review panel assessed OpenAI's planned transition away from its original non-profit structure as weakening alignment to its safety-focused mission. xAI is also registered as a Public Benefit Corporation, but has not yet demonstrated how that structure translates into meaningful safety governance.

The review panel weighted advocacy efforts related to AI regulations as a significant factor in their assessments. Panel members noted Anthropic's partial endorsement of SB 1047 and its opposition to a federal preemption of state-level regulation. x.AI's CEO was acknowledged for publicly supporting SB 1047. In contrast, experts reduced grades for OpenAI, Google DeepMind, and Meta for lobbying against key AI safety regulations, including the EU AI Act, SB 1047, and the RAISE Act.

The review panel identified robust public whistleblowing policies as an industry-wide gap, with panel members noting that OpenAI was the only company to publish its whistleblowing policy—a standard the panel considered a

basic expectation in safety-critical industries. While this transparency was seen as a positive step that other firms should emulate, OpenAI has also drawn criticism for its previous use of restrictive non-disclosure agreements tied to the vested equity of former employees. Experts flagged Google DeepMind and Meta for past incidents involving retaliation against employees who raised concerns, which panel members assessed as potentially having a chilling effect on safety culture.

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.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Domain Grade	A-	B	B	C+	D	D	F
Score	3.7	3.0	3.0	2.3	1.0	1.0	0

Indicator overview

Technical Specifications
- System Prompt Transparency
- Behavior Specification Transparency

Voluntary Cooperation
- G7 Hiroshima AI Process Reporting
- FLI AI Safety Index Survey Engagement

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

Compared to last year, OpenAI stood out for its engagement with the AI Safety Index survey. The firm's detailed model specification and regular disclosure of identified instances of malicious misuse were positively highlighted by reviewers.

Risk communication was found to diverge sharply between firms. Reviewers recognized Anthropic's proactive stance in informing policymakers and the public about critical risks, while Meta lost scores because its leadership publicly downplays extreme risks.

System prompt secrecy was found to still be the norm for proprietary models, with only Anthropic and xAI receiving credit for exposing the texts that steer their models.

Reviewers noted that incident reporting frameworks are largely absent, with none of the seven companies providing a concrete, public process for notifying governments about critical incidents. It was noted that this absence could undermine collective learning, and slow government responses to emerging threats.

xAI, ZhipuAI, and OpenAI received credit for submitting responses to the AI Safety Index survey. Regarding voluntary cooperation with the G7 Hiroshima AI Reporting Process, Meta and xAI stood out to reviewers as the only firms based in G7 countries that did not submit documentation on their risk management system.

5 Conclusions

The 2025 FLI AI Safety Index reveals an industry trust crisis: despite growing international consensus on AI risks and mounting evidence of rapid capability advances, experts warn that the gap between technological ambition and safety preparedness is widening. With no company achieving a grade higher than a C+ overall, reviewers expressed doubts that the industry's self-regulatory approaches will prove sufficient to address the magnitude of the risks.

The report's findings paint a troubling picture: Companies are racing toward artificial general intelligence and predict they will achieve superhuman performance within this decade. Yet as one reviewer noted, *"none of the companies has anything like a coherent, actionable plan"* for controlling such systems. While some firms invested in meaningful evaluations of dangerous capabilities despite fierce competitive pressure, reviewers still identified glaring methodological shortcomings across the industry. One expert warned,

> *"I have very low confidence that dangerous capabilities are being detected in time to prevent significant harm. Minimal overall investment in external 3rd party evaluations decreases my confidence further."*

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.

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.

You can skip between the domains by selecting one from the list below:

Domains

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

Internal Testing

Indicator

Dangerous Capability Evaluations

Definition & Scope

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, AI R&D facilitation, and behaviors associated with goal misalignment or deception. Evidence is drawn from model cards, including published results and detailed testing methodologies. The focus is on external deployments, as there is insufficient transparency on internal deployments.

Evaluation Guidance

Transparency Classification:

- Low detail: Only states that evaluations were conducted without naming specific tests or explaining methodologies.
- Moderate detail: Brief explanations of specific evaluations (and methodology).
- High details: Extensive explanations of individual evaluations and methodology.

Notes on AI regulation in China:

Under the *Interim Measures for the Management of Generative AI Services* (Aug 2023), every public-facing (B2B exemption) GenAI service must pass a government-supervised security assessment and complete an algorithm filing (算法备案); regulators may run their tests during this process.

Providers rely on the draft national standard "Basic Security Requirements for Generative AI Service" (TC260)

to prepare. The draft contains a list of 29 security risks in five buckets, yet, while an earlier draft version from Feb 2024 mentioned frontier risks, the current version does not [China Talk, 2024; China Law Translate].

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.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model	Claude 4 Opus	DeepSeek R1	Gemini 2.5 Pro	Llama 4 Maverick	o3	Grok 3	GLM-4
Bio + Chem	Yes ('CBRN') - High level of detail - Quantitative results with human & AI baselines - Safety framework classification - 10+ evaluations reported Evaluations include: Bioweapons acquisition uplift trial, Expert red-teaming (Deloitte), Long-form virology tasks, Multimodal virology (VCT), Bioweapons knowledge questions, DNA Synthesis Screening Evasion, LAB-Bench subset, Creative biology, Short-horizon computational biology, ASL-4 expert red-teaming System Card: pages 88-103	None	Yes ('CBRN') - Moderate level of detail - Quantitative results with AI (& human) baselines - Safety framework classification (CBRN uplift) Multiple-choice benchmarks, open-ended qualitative assessments led by domain experts across the biological, radiological, and nuclear domains. Three public benchmarks reported: SecureBio VMQA, FutureHouse LAB-Bench, Weapons of Mass Destruction Proxy. Model Card: pages 9-11	Yes ('CBRNE') - but no details provided - Minimal details - No quantitative results Reports expert-designed targeted evaluations and red-teaming without giving details [Meta]	Yes ('Bio') - Moderate level of detail - Quantitative results with human & AI baselines - Safety framework classification (Bio&Chem). Evaluations: Long-form biorisk questions, Multimodal troubleshooting virology, ProtocolQA Open-Ended Tacit knowledge, and troubleshooting Model Card: pages 12-15	None	None
Cyber offense	Yes ('Cybersecurity') - High level of detail - Quantitative results - Tracked in RSP, but no formal threshold Evaluations: Web Exploitation (15 CTFs), Cryptography (22 CTFs), Exploitation (9 CTFs), Reverse Engineering (8 CTFs), Network (4 CTFs), Cyber-harness network (3 ranges), Cybench (39 challenges) System Card: pages 116-122	None	Yes ('Cybersecurity') - High level of detail - Quantitative results with AI baselines - Safety framework classification (Cyber uplift + Cyber Autonomy) - Open-sourced evaluation suite 1) Previously published evaluation suite including In-house CTF (13), Hack The Box (13), Vulnerability detection (3) [arXiv, 2024]. 2) 50 additional challenges across four categories following their newly published framework: Reconnaissance, Tool development, Tool usage, Operational security [arXiv, 2025]. Model Card: pages 11-13	Yes ('Cyber attack enablement') - Minimal details - No quantitative results Card reports "evaluating the capabilities of Llama 4 to automate cyberattacks, identify and exploit security vulnerabilities, and automate harmful workflows". Does not give more details. [Meta]	Yes ('Cybersecurity') - Moderate level of detail - Quantitative results with AI baselines - Safety framework classification (Cybersecurity). Model card (p.15) 1) Two scenarios from the "Cyber Range" evaluation for conducting fully end-to-end cyber operations in a realistic, emulated network. 2) 100 capture the flag challenges across three difficulty levels (high school, collegiate, professional). Categories: Web Application Exploitation, Reverse Engineering, Binary and Network Exploitation, Cryptography, misc. Model Card: pages 15-22	None	None

Table continues on next page

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
AI R&D	Yes ('Autonomy') - High level of detail - Quantitative results with expert baselines - Safety framework classification (Autonomy) Evaluations: SWE-bench Verified (hard subset), METR data deduplication, Internal AI research suite 1 (6 tasks), Internal AI research suite 2, Internal model use survey System Card: pages 101-113	None	Yes ('ML R&D') - High level of detail - Quantitative results with human and AI baselines - Safety framework classification (ML R&D uplift) Used open-source Research Engineering Benchmark (RE-Bench, Wijk et al., 2024) https://arxiv.org/abs/2411.15114 Model Card: pages 14-16	None	Yes ('AI Self-improvement') - Moderate to High level of detail - Quantitative results with AI baselines - Safety framework classification (AI Self-improvement) - Additional external evaluation from METR on autonomy and AI R&D [METR] Five evaluations: OpenAI Research Engineer Interviews, SWE-bench Verified, OpenAI PRs, SWE-Lancer, PaperBench. Model Card: pages 22-28	None	None
Scheming capabilities & Misalignment	Yes ('Comprehensive Alignment Assessment') - High level of detail - First comprehensive alignment assessment in the industry (30 pages) - Additional external testing by Apollo Research Comprehensive Alignment Assessment covers: systematic deception, hidden goals, self-preservation, sandbagging, situational awareness, sycophancy, reasoning faithfulness, jailbreak susceptibility, values, high-agency behavior, subtle sabotage, Excessive compliance with harmful system prompts, pro-AI bias, and Reward seeking Apollo Research's evaluations are similar to a recent technical paper: "in-context scheming reasoning" [arXiv, 2024]. System Card: pages 19-48	None	Yes (Focuses on capabilities but not propensities) ('Deceptive alignment') - high level of detail - safety framework classification ("Instrumental Reasoning") - Evaluations open-sourced Five evaluations on ability to circumvent oversight (stealth), ("cover_your_tracks, decision_sabotage, strategic_ rule_breaking, oversight_pattern, classifier_evasion") and seven evaluations for measuring a model's ability to instrumentally reason about itself and its environment (situational awareness). Details published in a technical paper [arXiv, 2025]. Model Card: pages 15-16	None	Yes ('Deception / Scheming'), only external evaluations by Apollo Research - High level of detail - quantitative results with human and AI baselines Evaluations: strategic deception, in-context scheming, reasoning, and sabotage. Evaluations similar to recent technical paper: "in-context scheming reasoning" [arXiv, 2024]. Model Card: pages 10+30	None	None

Table continues on next page

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Transparency Overview	Model Card Length: 122 pages (Opus + Sonnet) **Safety Evaluations:** - 10.5 pages (p 11-21) **Frontier Risk Evaluations:** - 36 pages (p. 87-122) **External Evaluations:** - 2 pages (p. 30-31, 122) **Other:** 1) Comprehensive Alignment Assessment: 29 pages (p. 22-51) [Anthropic, 2025] 2) AI Safety Level 3 Deployment Safeguards Report 25 pages Content: Claude 4 Opus was classified as requiring AI Safety Level 3 (ASL-3) under their Responsible Scaling Policy, indicating it could potentially assist with CBRN weapons development. The relevant safeguards report (separate from the model card) outlines the core threat model, details the implemented safeguards, and provides evidence demonstrating their effectiveness [Anthropic, 2025].	Technical report length: 22 pages No content on safety evaluations [arXiv, 2025]	Model Card Length: 17 pages **Safety Evaluations:** - 2 pages (p. 5-7) **Frontier Risk Evaluations:** - 8 pages (p. 8-16) **External Evaluations:** - 0.5 pages (p. 10) **Linked Resources:** Additional results in technical paper: 'Evaluating Frontier Models for Stealth and Situational Awareness' (45 pages) [arXiv, 2025] **Other:** Announces: "detailed technical report will be published once per model family's release, with the next technical report releasing after the 2.5 series is made generally available." (Google considers the current release to be a "Preview") [Google, 2025].	Model Card Length: 14.5 pages (browser print format of website) **Safety Evaluations:** - 2 pages (p. 10-12) **Frontier risk evaluations:** - 1 page (p. 13-14) [Huggingface]	Model Card Length: 31.5 pages (o3 +o4 mini) **Safety Evaluations:** - 7 pages (p. 2-8) **Frontier Risk Evaluations:** - 16 pages (p. 11-27) **External Evaluations:** - 5 pages (p. 8-11, 30-32) **Other:** OpenAI's Safety Evaluations Hub webpage provides an ongoing overview of safety test results regarding harmful content, jailbreaks, hallucinations, and instruction hierarchy compliance. It currently shares updated evaluation results across 9 different AI models. [OpenAI, 2025; OpenAI, 2025]	No relevant model card found. The announcement post does not report safety evaluations. [xAI, 2025]	Technical report length: 12.5 pages Safety Evaluations: - 1 page (p. 12) [arXiv, 2024]

Sources

- Forum, Frontier Model. "Issue Brief: Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations." Frontier Model Forum, 14 Jan. 2025

Indicator

Elicitation for Dangerous Capability Evaluations

Definition & Scope

Assesses the extent to which a company discloses its elicitation strategy in its most recent dangerous-capability evaluations.* We record whether the company is transparent about:

1. Parallelization settings – e.g., *pass@ *n* and *best-of-n* sampling parameters (especially relevant for AI-R&D and cyber-security tasks).
2. Tooling – any use of internet access, code interpreters, agentic scaffolds, or relevant tools that can amplify model performance.
3. Model variants – the exact model checkpoints tested, including "helpful-only" variants and any domain- or task-specific fine-tuning.

Why This Matters

It has been demonstrated that small improvements in elicitation methodology can dramatically increase scores on evaluation benchmarks. Naive elicitation strategies cause significant underreporting of risk profiles, potentially missing dangerous capabilities that sophisticated actors could unlock. Companies thus need to implement comprehensive elicitation methodologies to better approximate an AI model's true capabilities, not just its default behavior. This should include task-specific fine-tuning in domains like bio-risk, especially if model weights will be made generally available, but also be general, as model weights might be stolen or leaked. A structured, transparent, and well-resourced approach to capability elicitation demonstrates a genuine commitment to risk discovery.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Parallel test-time compute & tooling	Mentions specific tools on tools and parallel computing approaches for several cyber evaluations. For cyber CTFs, pass@30 is reported. Bio-section: - "for automated evaluations, our models have access to various tools and agentic harnesses (software setups that provide them with extra tools to complete tasks)" - Some evaluations comment on the parallel test time compute approach, e.g., pass@5 for longform virology	No tests reported	None mentioned	None mentioned	Mentions specific tools on tools and parallel computing approaches for several cyber and self-improvement evaluations. For cyber CTFs, pass@12 is reported, for self-improvement, often pass@1. Multiple choice bio-risk questions were reported as consensus@32.	No tests reported	No tests reported
Model versions & Domain / Task-specific fine-tuning	- Tested helpful-only model without safety mitigations. - No mention of domain/task-specific fine-tuning. System Card: page 8	No tests reported	None mentioned	None mentioned	- Tested helpful-only model without safety mitigations. - No domain/task-specific fine-tuning reported System Card: page 13	No tests reported	No tests reported

Sources

- Adler, Steven. "AI Companies Should Be Safety-testing the Most Capable Versions of Their Models." Steven Adler's Substack, 26 Mar. 2025
- Metr. "Guidelines for Capability Elicitation." METR's Autonomy Evaluation Resources, 13 Mar. 2024Indicator

Indicator

Human Uplift Trials

Definition & Scope

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.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Yes (1) **Bioweapons Acquisition Uplift Trial:** - Methodology: Controlled trial with groups of 8-10 participants given up to 2 days to draft a comprehensive bioweapons acquisition plan - Groups: Control group: Only basic internet resources Model-assisted group: Additional access to Claude with safeguards removed - Participants: Contracted from SepalAI and Mercor - Grading: By Deloitte using a detailed rubric, assessing key steps of the acquisition pathway System Card: page 92-93	None	None	None	None	None	None

External Testing

Indicator

Independent Review of Safety Evaluations

Definition & Scope

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 underreport alarming results or select lenient testing conditions that avoid costly deployment delays. Regulators, investors, and the public 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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
None	None	None	None	None	None	None

Sources

- "Key Components of an RSP." METR, 26 Sept. 2023
- Homewood, Aidan, et al. "Third-party compliance reviews for frontier AI safety frameworks."

Indicator

Pre-deployment External Safety Testing

Definition & Scope

This indicator evaluates whether companies facilitate independent third-party safety assessments prior to releasing frontier models. It excludes collaborative testing arrangements and focuses solely on unaffiliated evaluators. 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

Independent evaluations are essential for verifying safety claims and uncovering risks that internal teams may miss, perhaps due to misaligned incentives or bias. Providing evaluators with substantial access—and ensuring their ability to publish freely—reflects a company's commitment to transparent, evidence-based governance.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
UK AI Security Institute & US Center for AI Standards and Innovation (CAISI) Scope: catastrophic risks in CBRN, cybersecurity, and autonomous capabilities Notice: Institutes will receive a minimally redacted copy of the internal capabilities report Details: No details **Apollo Research** Scope: Propensities and capabilities related to sabotage Methodology: "following methods from their prior work in Balesni et al. [arXiv, 2024], Meinke, A., et al. (2024) [arXiv, 2025]. Model version: Early snapshot Details: One-page summary, results worded by Apollo. **Notes:** Anthropic's transparency hub states, "external evaluations use API access with 'zero data retention' settings to prevent content storage." [Anthropic, 2024].	None reported	Model card reports leveraging "third-party external testers" for CBRN risks, but does not disclose any details.	None reported	**U.S. AI Safety Institute** Scope: Cyber and biological capabilities evaluations Model versions: Early checkpoints + final launch candidate models of both o3 and o4-mini Details: No details **U.K. AI Security Institute** Scope: Cyber, chemical, and biological autonomy capabilities, and an early version of safeguards Model versions: Early checkpoints + final launch candidate models Details: No details **METR (Model Evaluations and Threat Research)** Duration: 15 days Model versions: Earlier checkpoints of o4-mini and o3 Scope: Autonomous capabilities and reward hacking. Details: METR published a paper on methods [arXiv, 2025] and an evaluation report here [METR]. **Apollo Research** Scope: In-context scheming and strategic deception Model versions: Early checkpoints + final launch candidate models of both o3 and o4-mini Methodology: Similar to Apollo's technical paper on o1 evaluation. Meinke, A., et al. (2024). Frontier models are capable of in-context scheming [arXiv, 2025]. Details: <1 page summary of results + 3 pages with quantitative results in appendix. **Pattern Labs** Scope: Cybersecurity - three types of cyber offensive challenges: 1) Evasion, 2) Network Attack Simulation, 3) Vulnerability Discovery and Exploitation. Model versions: Early checkpoints + final launch candidate models of both o3 and o4-mini Details: <1 page summary of results System Card: page 9 **Additional details from the index survey** [Response]: (Questions reference this specific release) Q17: "In some instances, we paid private consultants for their work, but payment is not conditioned on the content of their findings." Q18: Highest level of access granted to any external evaluators: " 'Helpful-only' or base model API (no harmlessness fine-tuning and no filters)" Q19: Longest access period for any external evaluator: Between 2-3 weeks Q20: Publication arrangements: 1) Evaluators are under NDA, and publications require prior approval from OpenAI. 2) OpenAI publishes mutually agreed-upon excerpts of reports, with OpenAI having the final say on what content goes in System Cards.	None reported	None reported

Table continues on next page

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
				Q21: Query-rate or volume restrictions: 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). OpenAI notes that rates can sometimes depend on technical feasibility. Q22: Logging and retaining model interactions: "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)."		

Sources

- Che, Zora, et al. "Model Manipulation Attacks Enable More Rigorous Evaluations of LLM Capabilities." Neurips Safe Generative AI Workshop 2024.

Indicator

Bug Bounties for Model Vulnerabilities

Definition & Scope

This indicator assesses the presence and design of structured incentive programs—such as bug bounties or red-teaming initiatives—that encourage responsible disclosure of safety vulnerabilities in AI models. It focuses exclusively on programs addressing model behavior, excluding conventional cybersecurity initiatives due to insufficient public reporting. Evidence includes the scope of eligible issues, compensation levels, response timelines, and public availability of program documentation.

Why This Matters

Structured disclosure programs with financial incentives harness external expertise to identify model vulnerabilities before they are exploited in deployment. Investments in such programs indicate a proactive attitude toward risk identification.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Bug bounty on universal jailbreaks - Opened applications for early access testing of new safety mitigations. - Started May 2025 (last iteration ran August 2024) [Anthropic, 2024] - Up to $25,000 for verified universal jailbreak attacks that could expose vulnerabilities in critical, high-risk domains - Still accepting applications [Anthropic, 2025]	None	**Abuse Vulnerability Reward Program:** Accepts certain abuse-related discoveries: - Prompt Attacks - Training Data Extraction - Manipulating Models - Adversarial Perturbation - Model Theft (excludes jailbreaks) [Google]	Bounty programs are restricted to privacy or security issues, like extracting training data through tactics like model inversion or extraction attacks. [Meta]	**Early access for safety testing (December 2024)** One-off programs allowed safety researchers to apply for early access to frontier models to help surface novel risks. No payments announced. [OpenAI, 2024]	None	None

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Assessment methods provide very high confidence that dangerous capabilities would be detected in time to prevent significant harm
- **B** Assessment practices provide high confidence in detecting dangerous capabilities
- **C** Assessment approach provides moderate confidence with concerning gaps
- **D** Limited assessments provide low confidence in risk detection
- **F** No credible basis for detecting dangerous capabilities

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

Companies were evaluated on how effectively their models mitigate current harms, with a focus on safety benchmark performance, robustness against adversarial attacks, watermarking of AI-generated content, and the treatment of user data.

Table of Contents

Model Safety / Trustworthiness

Indicator

Stanford's HELM Safety Benchmark

Definition & Scope

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 (GPT-4o and Llama 3.1 405B) for the remaining benchmarks.

Why This Matters

HELM Safety offers a standardized evaluation of core safety risks. HELM enables consistent comparison across models. Strong performance signals effective safeguards against harmful outputs, including violence, fraud, discrimination, and other high-impact risks that affect millions of users.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model	Claude 4 Opus	DeepSeek R1	Gemini 2.5 Pro (03-25 preview)	Llama 4 Maverick	o3	Grok 3 Beta	n/a **(Model not evaluated by external benchmark)**
Average score (max score = 1)	0.97	0.87	0.91	0.91	0.98	0.86	
HarmBench	0.92	0.47	0.65	0.66	0.98	0.45	
SimpleSafetyTests	1.00	0.98	0.97	0.99	0.99	0.97	
BBQ accuracy	0.97	0.97	0.96	0.93	0.98	0.94	
Anthropic Red Team	0.99	0.96	1.00	0.98	0.98	0.96	
XSTest	0.97	0.95	0.99	0.97	0.97	0.96	
Retrieved	12 June 2025						
Release	Release v1.8.0						

Sources

- Farzaan, et al. "HELM Safety: Towards Standardized Safety Evaluations of Language Models." Stanford Center for Research on Foundation Models, 8 Nov. 2024. Accessed 14 Jun. 2025.

Indicator

Stanford's HELM AIR Benchmark

Definition & Scope

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 abstract or theoretical benchmarks, it directly reflects the kinds of standards developers will be expected to meet. Strong performance indicates that AI models comply with normative expectations around responsible AI deployment.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Pro	Llama 4 Maverick	o3	Grok 3 Beta	n/a **(Model not evaluated)**
Average score (max score = 1)	0.82	0.53	0.74	0.69	0.85	0.51	
System & Operational Risks	0.69	0.28	0.62	0.55	0.80	0.37	
Content Safety Risk	0.84	0.57	0.73	0.78	0.78	0.49	
Societal Risk	0.87	0.51	0.74	0.72	0.94	0.37	
Legal & Rights-related Risks	0.90	0.59	0.83	0.79	0.92	0.58	
Retrieved	28 May 2025						
Release	Release v1.11.0 (2025-04-23)						

Sources

- Zeng, Yi, et al. "Air-bench 2024: A safety benchmark based on risk categories from regulations and policies." 2024. Accessed 14 Jun. 2025.
- Website: Stanford's AIR-Bench 2024.

Indicator

TrustLLM Benchmark

Definition & Scope

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 provides a comprehensive evaluation of model reliability across multiple critical dimensions beyond traditional capability metrics. Strong performance across truthfulness, privacy protection, and ethical reasoning indicates that the companies have invested in aligning their models to be harmless and helpful, and not to cause unintended harm.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model Tested	Claude 4 Opus (extended thinking)	DeepSeek R1	Gemini 2.5 Pro	Llama 4 Maverick	o3	Grok 3 Beta	GLM 4 (32B)
Total score (max score = 1)	0.64	0.62	0.62	0.60	0.61	0.61	0.64
Truthfulness	0.54	0.62	0.57	0.53	0.43	0.59	0.57
Safety	0.65	0.61	0.62	0.60	0.65	0.60	0.63
Fairness	0.43	0.39	0.39	0.45	0.42	0.38	0.61
Privacy	0.60	0.58	0.57	0.53	0.57	0.55	0.52
Ethics	0.90	0.89	0.88	0.84	0.85	0.86	0.82
Robustness	0.67	0.62	0.64	0.61	0.66	0.65	0.67
Resources	Code & Generations, Full Score Breakdown						
Release	v0.3.0						

Sources

- Sun, Lichao, et al. "Trustllm: Trustworthiness in large language models." arXiv preprint arXiv:2401.05561 3 (2024).
- Website: TrustLLM

Robustness

Indicator

Gray Swan Arena: UK AISI Agent Red-Teaming Challenge

Definition & Scope

This indicator measures resistance to adversarial attacks based on the U.K. AISI Agent Red-Teaming Challenge conducted on the Gray Swan Arena platform. The challenge is the largest public evaluation to date of safety in agentic LLMs, testing how well AI agents can withstand attempts to manipulate or subvert their behavior. Participants used both direct and indirect red-teaming techniques—strategies designed to trick or exploit the model—to identify vulnerabilities across five core behavior categories, including Confidentiality Breaches and Instruction Hierarchy Violations. Performance is measured using the Attack Success Rate (ASR), calculated as Total Breaks divided by Total Chats, offering a concrete metric of model robustness under real-world adversarial pressure.

Why This Matters

Agent red-teaming resistance measures real-world robustness against sophisticated attacks that could compromise AI systems in deployment. Models with lower attack success rates demonstrate stronger defenses against attempts to violate safety policies, extract confidential information, or manipulate agent behavior. This competitive environment with expert red-teamers provides a more realistic assessment than academic benchmarks, revealing which companies have invested in hardening their systems against adversarial exploitation.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model Tested	Claude 3.7 Sonnet Thinking	n/a (Model not evaluated)	n/a (Model not evaluated)	Llama 4 Maverick	o3	Grok 3 Beta	n/a (Model not evaluated)
Attack Success Rate	1.45%			5.90%	2.46%	4.14%	
Retrieved	28 May 2025						

Sources

- Gray Swan AI. "UK AISI × Gray Swan Agent Red-Teaming Challenge: Results Snapshot." Gray Swan News, 2024.
- Website: Agent Red-Teaming Leaderboard

Indicator

Cisco Security Risk Evaluation

Definition & Scope

This indicator presents the results of a security risk assessment of frontier reasoning models, conducted by researchers from Cisco's Robust Intelligence team and the University of Pennsylvania. The experiments evaluated how resistant these models are to automated jailbreaking attacks—techniques designed to bypass safety systems and elicit harmful outputs. Researchers used 50 randomly selected prompts from the HarmBench dataset, covering six harm categories: cybercrime, misinformation, illegal activities, general harm, harassment, and chemical/biological weapons. The main metric reported is the Attack Success Rate (ASR), reflecting the percentage of harmful prompts for which a successful jailbreak was achieved. This provides a standardized way to compare the strength of safety guardrails across different models.

Why This Matters

Algorithmic jailbreaking resistance is a key measure of the robustness of safety guardrails. Models with high attack success rates are "highly susceptible to algorithmic jailbreaking and potential misuse," creating significant risks when deployed at scale.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model Tested	Claude 3.5 Sonnet	DeepSeek R1	Gemini 1.5 Pro	Llama 3.1 405B	GPT-4o | o1-preview	n/a (Model not evaluated)	n/a (Model not evaluated)
Attack Success Rate	36%	100%	64%	96%	86% | 26%		

Sources

- Kassianik, Paul. "Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models." Cisco Security Blog, January 31, 2025.

Indicator

Protecting Safeguards from Fine-tuning

Definition & Scope

This indicator evaluates whether companies implement safeguards to prevent the removal of safety measures through fine-tuning. Evidence distinguishes between hosted supervised fine-tuning, where inference-time mitigations remain in place, and full weight release without tamper-resistant safeguards. Where no specific data on frontier models is reported, neither fine-tuning nor open weights are accessible.

Why This Matters

Companies that release full model weights enable malicious actors to strip all safety protections and create uncensored versions, while supervised fine-tuning helps maintain safety guardrails during customization.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
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]	Fully released weights of frontier models. No tamper-resistant safeguards. [DeepSeek, 2025]	Frontier model weights protected Released weights of non-frontier Gemma family, including Gemma 3 27B [Hugging Face, 2025]. No tamper-resistant safeguards. Enables supervised fine-tuning of Gemini 2.0 Flash via Vertex AI. Safety mitigations are in place. [Google, 2025].	Fully released weights of the frontier model Llama 4 Maverick. No tamper-resistant safeguards. [Meta AI, 2025]	Frontier model weights protected Provide supervised fine-tuning of GPT-4o [OpenAI, 2024] and RL fine-tuning for o4 mini [OpenAI, 2025]. Safety mitigations are in place.	Frontier model weights protected Fully released weights of non-frontier Grok 1. No tamper-resistant safeguards. [xAI, 2024]	Fully released weights of the frontier model GLM-4 Z1 32B. No tamper-resistant safeguards. [THUDM, 2025]

Sources

- Qi, Xiangyu, et al. "Fine-tuning aligned language models compromises safety, even when users do not intend to!." arXiv preprint arXiv:2310.03693 (2023).
- Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. "Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b." arXiv preprint arXiv:2310.20624 (2023).
- Tamirisa, Rishub, et al. "Tamper-resistant safeguards for open-weight llms." arXiv preprint arXiv:2408.00761 (2024).

Digital Responsibility

Indicator

Watermarking

Definition & Scope

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 deployment rather than research alone, 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 enables the detection of AI-generated content, helping combat misinformation and digital fraud. Companies that deploy robust watermarking systems—along with public detection tools—help to uphold transparency and accountability.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Text-based	None found	None found	Yes - the SynthID system uses particular token selection to introduce a pattern that marks a text as AI-generated [Google DeepMind]. This can be identified by an online detector, access currently limited [Google, 2025].	None found	Research-only watermark, not shipped [The Verge, 2024]	None found	None found

Table continues on next page

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Image-based	Claude does not generate images	None found	Yes (SynthID) [Google DeepMind]: pattern is embedded in images, can be identified by an online detector, access currently limited [Google, 2025]	Yes, but detection is restricted: Including invisible marks, detectable by Meta's own detector and partner platforms, they have not opened-sourced the model [Meta, 2024].	Uses the C2PA standard to flag the metadata of images generated by ChatGPT [OpenAI, 2025]. Such metadata is trivial to remove [Forbes, 2024].	None found	None found

Sources

- Zhao, Xuandong, et al. "SoK: Watermarking for AI-Generated Content." arXiv preprint arXiv:2411.18479 (2024).
- NIST. "Reducing Risks Posed by Synthetic Content." (2024).

Indicator

User Privacy

Definition & Scope

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 This Matters

Opt-in policies and private deployment options enable greater respect for user privacy, especially in sensitive fields such as healthcare, law, and government.

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Default training on user inputs	No, unless the user opts in explicitly or the conversation is flagged for violating our Usage Policy [Anthropic, 2025].	Yes [DeepSeek, 2025]	Yes, but not in the enterprise version [Google, 2025]	Yes [Meta]	Yes, but no training on enterprise data (from ChatGPT Team, Enterprise, or API Platform) [OpenAI, 2025]	Yes [Ars Technica, 2024]	Yes
Frontier model weights available for private hosting	No	Yes [Huggingface, 2025]	No, but less-powerful models are open-sourced [Huggingface, 2025]	Yes [Meta]	No	No, but less-powerful models are open-sourced [xAI]	Yes [THUDM]

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Exceptional safety; trivial issues only; no serious harm potential
- **B** Strong safety; rare moderate issues; serious harms well-controlled
- **C** Adequate safety; some moderate issues; serious harms mostly controlled
- **D** Inadequate safety; frequent issues; serious harms poorly controlled
- **F** Dangerous systems; pervasive issues; 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 comprehensive analysis supporting this section was conducted by the non-profit research organisation SaferAI.

Table of Contents

Overview

The section focuses on framework documents and excludes other documents such as model cards. The comprehensive analysis for the indicators in this domain was conducted SaferAI, a leading governance and research non-profit focused on AI risk management. The organisation works to incentivize responsible AI practices through policy recommendations, research, and innovative risk assessment tools. *Note: The assessment contains living scores that are updated on a continuous basis. We extracted the scores from June 24, 2025.*

Frameworks:

- Anthropic - Responsible Scaling Policy v2.2
- OpenAI - Preparedness Framework v2
- Google DeepMind - Frontier Safety Framework v2.0
- Meta - Frontier AI Framework v1.1
- xAI - Risk Management Framework (Draft)
- DeepSeek & Zhipu AI have not published a framework

	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
Overall score		36%	34%	22%	26%	19%
1. Risk Identification	25%	28%	27%	17%	33%	5%
2. Risk Analysis & Evaluation	25%	26%	34%	31%	36%	31%
3. Risk treatment	25%	40%	36%	23%	19%	18%
4 Risk Governance	25%	48%	38%	16%	15%	23%

Supporting references: Reviewers were provided with the full database of framework references and quotes supporting SaferAI's assessments of individual all sub-indicators. The data can be found on their project website.

Indicator

1. Risk Identification

Definition & Scope

This dimension captures the extent to which the company has addressed known risks in the literature and engaged in open-ended red teaming to uncover potential new threats. Moreover, this dimension examines if the AI company has leveraged a diverse range of risk identification techniques, including threat modeling when appropriate, to gain 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.

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
	1. Risk Identification	25%	28%	27%	17%	33%	5%
	1.1 Classification of Applicable Known Risks	**40%**	**30%**	**30%**	**18%**	**13%**	**13%**
C1	1.1.1 Risks from literature and taxonomies are well covered	50%	50%	50%	25%	25%	25%
C2	1.1.2 Exclusions are justified and documented	50%	10%	10%	10%	0%	0%
	1.2 Identification of Unknown Risks (Open-ended red teaming)	**20%**	**0%**	**3%**	**0%**	**0%**	**0%**
	1.2.1 Internal	70%	0%	3%	0%	0%	0%
C3	1.2.1.1 Adequate methodology (includes resources, time, and access to the model)	70%	0%	0%	0%	0%	0%
C4	1.2.1.2 Appropriate expertise to properly identify hazards	30%	0%	10%	0%	0%	0%
	1.2.2 Third parties	30%	0%	3%	0%	0%	0%
C5	1.2.2.1 Appropriate expertise to identify hazards	33%	0%	0%	0%	0%	0%
C6	1.2.2.2 Adequate resources/time/access to the model	33%	0%	10%	0%	0%	0%
C7	1.2.2.3 Commitment to non-interference with findings	34%	0%	0%	0%	0%	0%
	1.3 Risk Modeling	**40%**	**41%**	**36%**	**25%**	**69%**	**1%**
C8	1.3.2 The company uses risk models for all the risk domains identified, and the risk models are published (with potentially dangerous information redacted)	40%	50%	75%	50%	90%	0%
	1.3.1 Risk modeling methodology	40%	40%	10%	12%	58%	2%
C9	1.3.1.1 Methodology precisely defined	70%	50%	10%	10%	75%	0%
C10	1.3.1.2 Mechanism to incorporate red teaming findings	15%	10%	10%	10%	10%	0%
C11	1.3.1.3 Prioritization of severe and probable risks	15%	25%	10%	25%	25%	10%
C12	1.3.4 Third-party validation of risk models	20%	25%	10%	0%	50%	0%

Indicator

2. Risk Analysis and Evaluation

Definition & Scope

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 crossing specific KRI thresholds triggers corresponding KCI requirements.

Why This Matters

Without operationalizing risk tolerances into measurable metrics, companies cannot make consistent, 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.

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
	2. Risk Analysis and Evaluation	25%	26%	34%	31%	36%	31%
	2.1 Setting a Risk Tolerance	**35%**	**7%**	**16%**	**3%**	**24%**	**57%**
	2.1.1 Risk tolerance is defined	80%	8%	20%	3%	30%	71%
C13	2.1.1.1 Risk tolerance is at least qualitatively defined for most risks	33%	25%	50%	10%	90%	90%
C14	2.1.1.2 Risk tolerance is expressed fully quantitatively (cf. criterion above) or at least partly quantitatively as a combination of scenarios (qualitative) and probabilities (quantitative) for most risks	33%	0%	10%	0%	0%	75%
C15	2.1.1.3 Risk tolerance is expressed fully quantitatively as a product of severity (quantitative) and probability (quantitative) for most risks	33%	0%	0%	0%	0%	50%
	2.1.2 Process to define the tolerance	20%	0%	0%	0%	0%	0%
C16	2.1.2.1 AI developers engage in public consultations or seek guidance from regulators where available.	50%	0%	0%	0%	0%	0%
C17	2.1.2.2 Any significant deviations from risk tolerance norms established in other industries are justified and documented (e.g., cost-benefit analyses)	50%	0%	0%	0%	0%	0%
	2.2 Operationalizing Risk Tolerance	**65%**	**36%**	**44%**	**47%**	**43%**	**18%**
	2.2.1 Key Risk Indicators (KRI)	30%	51%	51%	51%	33%	33%
C18	2.2.1.1 KRI thresholds are at least qualitatively defined for most risks	45%	90%	90%	90%	50%	50%
C19	2.2.1.2 KRIs thresholds are quantitatively defined for most risks	45%	50%	25%	10%	0%	75%
C20	2.2.1.3 KRI also identifies and monitors changes in the level of risk in the external environment	10%	0%	0%	0%	10%	0%
	2.2.2 Key Control Indicators (KCI)	30%	31%	45%	38%	18%	14%
	2.2.2.1 Containment KCIs	35%	43%	5%	63%	38%	5%
C21	2.2.2.1.1 Most KRI thresholds have corresponding qualitative containment KCI thresholds	50%	75%	10%	75%	75%	10%
C22	2.2.2.1.2 Most KRI thresholds have corresponding quantitative containment KCI thresholds	50%	10%	0%	50%	0%	0%
	2.2.2.2 Deployment KCIs	35%	38%	45%	25%	13%	25%
C23	2.2.2.2.1 Most KRI thresholds have corresponding qualitative deployment KCI thresholds	50%	75%	90%	50%	25%	50%
C24	2.2.2.2.2 Most KRI thresholds have corresponding quantitative deployment KCI thresholds	50%	0%	0%	0%	0%	0%

Table continues on next page

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
C25	2.2.2.3 For advanced KRIs, Assurance processes KCI are defined	30%	10%	90%	25%	0%	10%
C26	2.2.3 Pairs of thresholds are grounded in risk modeling to show that risks remain below the tolerance	20%	10%	50%	90%	50%	10%
C27	2.2.4 Policy to put development on hold if the required KCI threshold cannot be achieved, until sufficient controls are implemented to meet the threshold	20%	50%	25%	10%	90%	10%

Indicator

3. Risk Treatment

Definition & Scope

This dimension captures the extent to which the company has implemented comprehensive risk mitigation strategies across three critical areas: containment measures that control access to AI models, deployment measures that prevent misuse and accidental harms, and assurance processes that provide affirmative evidence of safety. Furthermore, this dimension assesses whether the company maintains continuous monitoring of both Key Risk Indicators (KRIs) and Key Control Indicators (KCIs) 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.

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
	3. Risk treatment	25%	40%	36%	23%	19%	18%
	3.1 Implementing Mitigation Measures	**50%**	**24%**	**32%**	**19%**	**13%**	**18%**
	3.1.1 Containment measures	35%	25%	25%	0%	0%	0%
C28	3.1.1.1 Containment measures are precisely defined for all KCI thresholds	60%	25%	25%	0%	0%	0%
C29	3.1.1.2 Proof that containment measures are sufficient to meet the thresholds	40%	25%	10%	0%	0%	0%
C30	3.1.1.3 Strong third-party verification process to verify that the containment measures meet the threshold	formula*	25%	25%	0%	0%	0%
	3.1.2 Deployment measures	35%	40%	50%	35%	35%	50%
C31	3.1.2.1 Deployment measures are precisely defined for all KCI thresholds	60%	50%	50%	25%	25%	25%
C32	3.1.2.2 Proof that deployment measures are sufficient to meet the thresholds	40%	25%	50%	50%	50%	50%
C33	3.1.2.3 Strong third-party verification process to verify that the deployment measures meet the threshold	formula*	**10%**	**25%**	**0%**	**0%**	**50%**
	3.1.3 Assurance processes	30%	5%	20%	23%	2%	3%
C34	3.1.3.1 Credible plans towards the development of assurance properties	40%	10%	25%	25%	0%	10%
C35	3.1.3.2 Evidence that the assurance properties are enough to achieve their corresponding KCI thresholds	40%	0%	25%	25%	0%	0%
C36	3.1.3.3 The underlying assumptions that are essential for their effective implementation and success are clearly outlined	20%	10%	10%	25%	10%	0%

Table continues on next page

*Formula: 100% if greater than the weighted average of 3.1.1.1 and 3.1.1.2

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
	3.2 Continuous Monitoring and Comparing Results with Pre-determined Thresholds	50%	56%	40%	27%	26%	17%
	3.2.1 Monitoring of KRIs	50%	68%	39%	27%	26%	4%
C37	3.2.1.1 Justification that the elicitation methods used during the evaluations are comprehensive enough to match the elicitation efforts of potential threat actors	30%	75%	75%	25%	50%	0%
	3.2.1.2 Evaluation frequency	25%	100%	0%	25%	0%	0%
C38	3.2.1.2.1 Specification of evaluation frequency in terms of the relative variation of effective computing power used in training	50%	100%	0%	25%	10%	0%
C39	3.2.1.2.2 Specification of evaluation frequency in terms of fixed time intervals to account for post-training enhancements	50%	100%	0%	0%	0%	0%
C40	3.2.1.4 Description of how post-training enhancements are factored into capability assessments	15%	75%	75%	50%	50%	0%
C41	3.2.1.5 Replication of evaluations by third parties	15%	50%	25%	25%	25%	25%
C42	3.2.1.6 Vetting of protocols by third parties	15%	10%	10%	10%	0%	0%
	3.2.2 Monitoring of KCIs	40%	33%	33%	23%	20%	15%
C43	3.2.2.1 Detailed description of evaluation methodology and justification that KCI thresholds will not be crossed unnoticed	40%	25%	25%	50%	50%	0%
C44	3.2.2.2 Replication of evaluations by third parties	30%	50%	50%	0%	0%	50%
C45	3.2.2.3 Vetting of protocols by third parties	30%	25%	25%	10%	0%	0%
C46	3.2.3 Sharing of evaluation results with relevant stakeholders as appropriate	10%	90%	75%	50%	50%	90%

Indicator

4. Risk Governance

Definition & Scope

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, whether they publicly disclose their 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.

ID	Criteria	Weight	Anthropic	OpenAI	Google DeepMind	Meta	x.AI
	4 Risk Governance	25%	48%	38%	16%	15%	23%
	4.1 Decision-making	**25%**	**44%**	**28%**	**13%**	**30%**	**40%**
C47	4.1.1 The company has clearly defined risk owners for every key risk identified and tracked	25%	25%	10%	0%	10%	75%
C48	4.1.2 The company has a dedicated risk committee at the management level that meets regularly	25%	0%	0%	0%	25%	0%
C49	4.1.3 The company has defined protocols for how to make go/no-go decisions	25%	75%	50%	50%	75%	10%
C50	4.1.4 The company has defined escalation procedures in case of incidents	25%	75%	50%	0%	10%	75%
	4.2. Advisory and Challenge	20%	**35%**	**53%**	**28%**	**21%**	**4%**
C51	4.2.1 The company has an executive risk officer with sufficient resources	17%	75%	0%	0%	0%	0%
C52	4.2.2 The company has a committee advising management on decisions involving risk	17%	10%	90%	90%	25%	0%
C53	4.2.3 The company has an established system for tracking and monitoring risks	17%	50%	75%	50%	50%	25%
C54	4.2.4 The company has designed people who can advise and challenge management on decisions involving risk	17%	25%	75%	0%	25%	0%
C55	4.2.5 The company has an established system for aggregating risk data and reporting on risk to senior management and the Board	17%	50%	75%	25%	25%	0%
C56	4.2.6 The company has an established central risk function	17%	0%	0%	0%	0%	0%
	4.3 Audit	20%	50%	38%	5%	5%	25%
C57	4.3.1 The company has an internal audit function involved in AI governance	50%	25%	0%	0%	0%	0%
C58	4.3.2 The company involves external auditors	50%	75%	75%	10%	10%	50%
	4.4 Oversight	**20%**	**50%**	**45%**	**25%**	**0%**	**0%**
C59	4.4.1 The Board of Directors of the company has a committee that provides oversight over all decisions involving risk	50%	25%	90%	0%	0%	0%
C60	4.4.2 The company has other governing bodies outside of the Board of Directors that provide oversight over decisions	50%	75%	0%	50%	0%	0%
	4.5 Culture	10%	**63%**	**12%**	**7%**	**3%**	**50%**
C61	4.5.1 The company has a strong tone from the top	33%	50%	25%	10%	10%	25%
C62	4.5.2 The company has a strong risk culture	33%	50%	10%	10%	0%	50%
C63	4.5.3 The company has a strong speak-up culture	33%	90%	1%	0%	0%	75%
	4.6 Transparency	5%	**72%**	**53%**	**20%**	**33%**	**45%**
C64	4.6.1 The company reports externally on what its risks are	33%	50%	75%	25%	50%	75%
C65	4.6.2 The company reports externally on what its governance structure looks like	33%	75%	75%	25%	25%	10%
C66	4.6.3 The company shares information with industry peers and government bodies	33%	90%	10%	10%	25%	50%

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Framework virtually guarantees prevention of catastrophic risks
- **B** Framework prevents catastrophic risks with a high degree of confidence
- **C** Framework prevents catastrophic risks with a moderate degree of confidence
- **D** Framework prevents catastrophic risks with a low degree of confidence
- **F** Framework prevents catastrophic risks with a very low degree of confidence; or no framework exists

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

Indicator

Existential Safety Strategy

Definition & Scope

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-defence 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. Transparency in safety strategies enables accountability and allows policymakers, researchers, and the public to evaluate whether adequate precautions are being taken.

Anthropic	Quantitative safety plan	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**The Urgency of Interpretability (2025, ~5k words, strategy blog)** The CEO argues in a personal blog that mechanistic interpretability must advance rapidly to ensure safe deployment of transformative AI systems that could become a "country of geniuses in a datacenter" by 2026-2027. Amodei frames this as a "race between interpretability and model intelligence" and outlines recommendations for the AI community and governments. The blog also discusses the history of interpretability research and recent technical breakthroughs. **Key quotes:** ▪ "Anthropic is doubling down on interpretability, and we have a goal of getting to "interpretability can reliably detect most model problems" by 2027. ▪ "Our long-run aspiration is to be able to look at a state-of-the-art model and essentially do a 'brain scan': a checkup that has a high probability of identifying a wide range of issues, including tendencies to lie or deceive, power-seeking, flaws in jailbreaks, [...]. This would then be used in tandem with the various techniques for training and aligning models, [..]." **Putting up Bumpers (2025, ~5k words, research blog)** Anthropic alignment researcher Sam Bowman proposes an alignment approach for early AGI systems that prioritizes implementing and testing "many largely-independent lines of defense" to catch and correct misalignment through iterative testing. He highlights "alignment audits" [Anthropic, 2025] as the "Primary Bumper" to notice signs of misalignment like "generalized reward-tampering" [Anthropic, 2024] or "alignment-faking" [Anthropic, 2024]. **Key quotes:** ▪ "Even if we can't solve alignment, we can solve the problem of catching and fixing misalignment." ▪ "We believe that, even without further breakthroughs, this work can almost entirely mitigate the risk that we unwittingly put misaligned circa-human-expert-level agents in a position where they can cause severe harm." ▪ "This is not a costless choice: The Bumpers' worldview largely gives up on the ability to make highly-confident, principled arguments for safety, and it comes with real risks." ▪ "We are plausibly within a couple of years of developing models that could automate much of the work of AI R&D. This makes sabotage and sandbagging threat models... worth addressing soon." ▪ "Anthropic is committed to investing seriously in the kinds of measures described here, ... setting up a new team to productionize and professionalize the hands-on work of testing models for AGI-relevant forms of misalignment." **Responsible Scaling Policy (2023, v2.2 in 2025, ~10k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain. **Core Views on AI Safety (2023, ~6k words, strategy blog)** This blog post outlines Anthropic's AI safety philosophy and technical research portfolio. The document addresses existential risk scenarios, presenting a three-tier framework (optimistic, intermediate, pessimistic) for how difficult alignment might prove to be, with corresponding strategic responses for each scenario. It details six priority research areas: Mechanistic Interpretability, Scalable Oversight, Process-Oriented Learning, Understanding Generalization, Testing for Dangerous Failure Modes, and Evaluating Societal Impact. The post emphasizes empirical research and acknowledges fundamental uncertainty about which approaches will succeed. **Key quotes:** ▪ "Our goal is essentially to develop: 1) better techniques for making AI systems safer; 2) better ways of identifying how safe or unsafe AI systems are." ▪ "We aim to build detailed quantitative models of how these [dangerous] tendencies vary with scale so that we can anticipate the sudden emergence of dangerous failure modes in advance." ▪ In pessimistic scenarios where "AI safety may be unsolvable," Anthropic's role would be "to provide evidence that current safety techniques are insufficient and to push for halting AI progress to prevent catastrophic outcomes."

Table continues on next page

DeepSeek	Quantitative safety plan	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	*Based on searches of company websites, technical papers, and public communications, no relevant strategy documents were found.*
Google DeepMind	Quantitative safety plan	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**An Approach to Technical AGI Safety and Security (2025,~80k words, technical report/research agenda)** A detailed technical report by DeepMind's safety team explains their research agenda for a framework to prevent severe, civilization-scale harm from AGI, defined as systems roughly at the 99th-percentile of skilled adults. The document states that reaching AGI before 2030 is plausible. However, it makes clear that a plan for ASI and a strong recursive Self-improvement is beyond the scope of this paper. **Key sections:** 'Assumptions about AGI development' (p. 20-43), 'Risk areas' (p. 44-55), 'Addressing misuse' (p. 56-70), 'Addressing Misalignment' (p. 70-108). **Scope:** "We identify four areas of risk: misuse, misalignment, mistakes, and structural risks. Of these, we focus on technical approaches to misuse and misalignment." The framework bets on a defence-in-depth, combining model-centric alignment research with classic security engineering. The plan is deliberately incremental: succeed on today's frontier models, then re-validate at each capability level rather than solve all of AGI safety in one stroke. The report is explicit about significant problems that remain unsolved (scalable interpretability, quantitative safety-case metrics, red-team rigor), which the authors flag as research priorities. "Importantly, this is a roadmap rather than a solution, as there remain many open research problems to address." **Misalignment Strategy - "Two Lines of Defense":** First Line - Model-level mitigations: "[..] ensure the model does not pursue misaligned goals. We follow the standard ML playbook: (1) improve the ability to get good oversight (amplified oversight), (2) use this to improve the model (guiding model behavior), and (3) figure out where oversight is needed to train a model that will robustly do what we want (robust training)." ▪ Amplified oversight: Debate, critique, constitutional AI, recursive reward modeling ▪ Guiding model behavior: Preference learning and value alignment techniques ▪ Robust AI oversight: Active learning, adversarial training, uncertainty estimation Second Line - System-level controls (monitoring, access) that still catch harm if alignment fails, including using dedicated monitor AI systems to detect misaligned actions, with emphasis on the monitor knowing when it's uncertain and flagging actions for review. ▪ Security measures: Treating AI as an "untrusted insider," sandboxing, access control ▪ Safer design patterns: Correctability, bounded autonomy, externalized reasoning ▪ Interpretability: The paper's longest technical section, covering sparse autoencoders, circuit analysis, and mechanistic interpretability ▪ Alignment stress tests: Red-teaming alignment assumptions using "model organisms" **Misuse strategy:** 3 Steps: 1. Evaluate models for dangerous capability. 2. Mitigate with post-training tuning, capability suppression, monitoring, and weight/access security. 3. Red team to build misuse safety cases (inability & control). Safety-case methodologies: Advocates evidence-based deployment decisions: inability cases (model lacks capability) and control cases (misaligned behaviour will be caught). Future work sketches empirical and incentive-based cases for more capable systems. **AGI Safety and Alignment at Google DeepMind: A Summary of Recent Work (2024, 2k words, research blog)** This update from DeepMind's team focused on existential risk describes their three main research bets over the past 1.5 years: 1) amplified oversight for proper alignment signals, 2) frontier safety to assess catastrophic risk capabilities, and 3) Mechanistic interpretability as an enabler for both. The post provides detailed explanations of recent work in each area and its rationales. The team admits they are "revising our own high-level approach to technical AGI safety" because current bets "do not necessarily add up to a systematic way of addressing risk." They highlight fundamental gaps, noting that even perfect amplified oversight would be insufficient under distribution shift, requiring additional investments in adversarial training, uncertainty estimation, and monitoring through a control framework. Frontier Safety Framework v2 **(2024, v2 in 2025, 4k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.

Table continues on next page

Meta	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**Frontier AI Framework v.1.1 (2025, ~8k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain. **Open Source AI Is the Path Forward (2024, ~3k words, strategy blog)** 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. He argues that widely deployed AI systems enable larger actors to check malicious uses by smaller actors. It addresses both unintentional harms (including "truly catastrophic science fiction scenarios for humanity") and intentional misuse by bad actors. **Key quotes:** ▪ "I think it will be better to live in a world where AI is widely deployed so that larger actors can check the power of smaller bad actors."
OpenAI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**How we think about safety and alignment (2025, ~3k words, strategy blog).** This blog describes high-level principles that guide OpenAI's thinking and ties it to their safety practices. This document describes a shift from viewing AGI as a single transformative moment to seeing it as continuous progress. For every principle, the blog lays out how it will shape their focus and approach to new challenges and relates to already implemented interventions. Quote of the core principles: 1) "Embracing uncertainty: We treat safety as a science, learning from iterative deployment rather than just theoretical principles.' 2) "Defense in depth: We stack interventions to create safety through redundancy." 3) "Methods that scale: We seek out safety methods that become more effective as models become more capable." 4) "Human control: We work to develop AI that elevates humanity and promotes democratic ideals." 5) "Community effort: We view responsibility for advancing safety as a collective effort." **Planning for AGI and beyond (2023, 2k words)** This high-level blog outlines principles for managing AGI risks. The post emphasizes goals like ensuring AGI benefits are "widely and fairly shared" and advocates for deploying progressively more powerful systems to learn iteratively. It acknowledges the need for new alignment techniques, calls for a global conversation on governance and benefit-sharing, describes the benefits of OpenAI's non-profit structure, and raises the idea of a coordinated slowdown. **Key quotes:** ▪ "We will need to develop new alignment techniques as our models become more powerful (and tests to understand when our current techniques are failing). Our plan in the shorter term is to use AI to help humans evaluate the outputs of more complex models and monitor complex systems, and in the longer term to use AI to help us come up with new ideas for better alignment techniques." ▪ "As our systems get closer to AGI, we are becoming increasingly cautious with the creation and deployment of our models. Our decisions will require much more caution than society usually applies to new technologies, and more caution than many users would like. Some people in the AI field think the risks of AGI (and successor systems) are fictitious; we would be delighted if they turn out to be right, but we are going to operate as if these risks are existential." **Announcement of Superalignment team (2023, ~1k words, strategy blog)** Outlined an ambitious strategy to start a new team to build "a roughly human-level automated alignment researcher" that could use vast compute to iteratively align superintelligence. Note: This team was disbanded in 2024 after team leaders Leike and Sutskever left OpenAI [CNBC, 2024]. Preparedness Framework **(2023, v2 in 2025, ~10k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.

Table continues on next page

x.AI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	xAI Risk Management Framework **(Draft) (2025, ~2k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.
Zhipu AI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	*Based on searches of company websites, technical papers, and public communications, no official strategy documents were found.* Media report on superalignment initiative **(National Business Daily, 2024)** At the AWS China Summit (Shanghai, 29 May 2024) Zhipu AI's Chief Ecosystem Officer Liu Jiang said: "AGI will reach ordinary-human level within 5-10 years.". He announces that "Zhipu AI has already launched a 'Superalignment' initiative." The article explains superalignment as "ensuring a super-human-level AI system follows human values and goals."

Sources

- Bengio, Yoshua, et al. "The Singapore Consensus on Global AI Safety Research Priorities," 2025.

Indicator

Internal Monitoring and Control Interventions

Definition & Scope

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 model has longer time-horizon interactions with the external world. Robust monitoring and control measures 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, technically sophisticated control protocols indicate that companies are taking tail risks seriously and work on mitigations.

Anthropic	**Stated AI Control Plans**	No 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].
	Implemented safeguards	None found
DeepSeek	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
Google DeepMind	**Stated AI Control Plans**	Early sketch of a plan without technical details. **Details:** Relevant section of Frontier Safety Framework: "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 Framework 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." [Google DeepMind, 2025].
	Implemented safeguards	None found
Meta	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
OpenAI	**Stated AI Control Plans**	No public control plan found Voluntary commitment in a safety framework requires "safeguards meeting a high standard" if the relevant capability threshold is crossed. Exact safeguard not specified, but several are suggested. **Details:** 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 agents 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]
	Implemented safeguards	None found

Table continues on next page

x.AI	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
Zhipu AI	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found

Sources

- Stix, Ch. AI Behind Closed Doors: a Primer on The Governance of Internal Deployment. Apollo Research, 17 Apr. 2025
- Carlsmith, Joe. "New Report: 'Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power?'" Joe Carlsmith, 15 Nov. 2023,
- Greenblatt, Ryan, et al. "AI control: Improving safety despite intentional subversion." arXiv preprint arXiv:2312.06942 (2023).

Indicator

Technical AI Safety Research

Definition & Scope

This indicator tracks 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, and model cards. The full collection was created by Zach Stein-Perlman—numbers for DeepSeek, ZhipuAI, and xAI added by FLI.

Why This 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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Total	32	0	28	6	12	0	0
2025	9	-	4	1	3	Safety advisor Dan Hendrycks publishes research, but not formally for xAI	0
2024	11	-	11	5	7		0
2023	12	-	13	-	2		0

Sources

- Stein-Perlman, Zach. "Boosting Safety Research." AI Lab Watch. Accessed 16 Jun. 2025.

Indicator

Supporting External Safety Research

Definition & Scope

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 models, 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 enable the research community by providing deeper

access to frontier models, mentoring the next generation of research talent, and empowering 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 enable academics and independent researchers with limited financial resources to experiment on frontier models.

Providing Deep Model Access to Safety Researchers

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
AI Safety researcher Ryan Greenblatt from Redwood Research was recently given employee-level access. [LessWrong, 2023]	Frontier model weights are publicly available	Non-frontier model Gemma 3 model weights publicly available [Google, 2025]	Frontier model weights are publicly available	OpenAI offers a public RL fine-tuning API. [OpenAI]	Non-frontier model Grok-1 model weights are publicly available [xAI, 2024]	Frontier model weights are publicly available

Sources:

- Shevlane, Toby, et al. "Model Evaluation for Extreme Risks." arXiv, 24 May 2023
- Casper, Stephen, et al. "Black-Box Access Is Insufficient for Rigorous AI Audits." arXiv, 29 May 2024

Mentoring and Funding

Anthropic	**Mentoring:** They have their own Anthropic Fellows program and provide a high number of mentors for the independent research seminar program MATS. [Anthropic, 2024; MATS, 2025] **External Researcher Access Program (ongoing):** ▪ gives free API credit to safety/alignment researchers ▪ Standard usage policies apply ▪ $1000 in API Credits (sometimes more) [Anthropic, 2025] **Initiative for developing third-party model evaluations (Jul 2024):** One-off program to provide funding for a third-party to develop evaluations that can effectively measure advanced capabilities in AI models: "The approach is designed to enable you to distribute your evaluations to governments, researchers, and labs focused on AI safety." [Anthropic, 2024].
DeepSeek	None found
Google DeepMind	**Mentoring:** Provides a high number of mentors for the independent research seminar program MATS. [MATS, 2025; MATS]
Meta	None found
OpenAI	**Mentoring:** Currently provides one mentor for the independent research seminar program, MATS. **Researcher Access Program (back since February 2025):** ▪ gives free API credit to safety/alignment researchers ▪ Standard usage policies apply ▪ Up to $1,000 of API credits [OpenAI, 2025] **Superalignment Fast Grants (2023):** $10M to support technical research towards the alignment and safety of superhuman AI systems, including weak-to-strong generalization, interpretability, scalable oversight, and more [OpenAI, 2023].
x.AI	None found
Zhipu AI	None found

Sources

- "Shevlane, Toby, et al. "Model Evaluation for Extreme Risks." arXiv, 24 May 2023
- Casper, Stephen, et al. "Black-Box Access Is Insufficient for Rigorous AI Audits." arXiv, 29 May 2024.

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Strategy provides strong quantitative guarantees against catastrophic risks from superintelligent AI
- **B** Strategy very likely to prevent catastrophic risks from superintelligent AI
- **C** Strategy likely to prevent catastrophic risks from superintelligent AI
- **D** Strategy may prevent catastrophic risks from superintelligent AI
- **F** Strategy unlikely to prevent catastrophic risks from superintelligent AI, or Strategy increases 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	

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. Indicators examine whistleblowing systems, legal structures, and advocacy on AI regulations.

Table of Contents

Indicator

Lobbying on AI Safety Regulations

Definition & Scope

This indicator documents a company's efforts to influence laws and regulations relevant to AI safety. It compiles publicly available evidence on direct policy positions—such as written statements, consultation responses, testimony, blog posts, and reputable media coverage—and records indirect engagement through membership in trade associations or coalitions that lobby on key safety rules.

Why This 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.

Sub-Indicator	Anthropic	DeepSeek	Google	Meta	OpenAI	x.AI	Zhipu AI
EU AI Act	No publicly available information found	No publicly available information found	Between 2022 and 2023, DeepMind lobbied EU institutions not to classify general-purpose AI and foundational models as "high-risk" technologies, a designation that would have triggered stricter safety obligations [Tranberg, 2023; TIME, 2023]. Google argued that the classification would hinder innovation, and regulations should attach further down the value chain [POLITICO, 2025; Data Ethics, 2023].	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 [Open Letter, 2023]. The company argued that strict obligations could hinder innovation and pushed for open-source models to be excluded from high-risk classification [Politico, 2025]. Chief AI Scientist Yann LeCun also criticized the EU's approach as overly restrictive [X, 2023].	In 2023, OpenAI lobbied EU officials to weaken parts of the AI Act, arguing that foundation models like GPT-4 should not face strict obligations unless adapted for specific uses [TIME, 2023; The Verge, 2023]. The company also pushed to delay transparency requirements and limit liability for general-purpose models.	No publicly available information found	No publicly available information found
Comprehensive US State-Level Regulation US	**California's SB 1047** In 2024, Anthropic initially raised concerns about California's SB 1047, influencing changes to the bill that softened key provisions [TechCrunch, 2024]. While the company opposed aspects of the original text, CEO Dario Amodei later expressed cautious support, stating in a letter to the governor that the bill's "benefits likely outweigh its costs" [Sanity.io, 2024]. Anthropic's involvement shaped the final version of the legislation [Vox, 2024].	No publicly available information found	**California's SB 1047** In 2024, Google DeepMind opposed California's SB 1047, arguing that its safety rules would burden developers and stifle innovation. The company warned that requirements like pre-deployment evaluations and state oversight could fragment regulation and urged alignment with federal efforts instead [DocumentCloud, 2024; Carnegie Endowment, 2024]. **Responsible AI Safety and Education (RAISE) Act** In 2025, industry groups with ties to Google DeepMind—including Tech:NYC and the Computer & Communications Industry Association (CCIA)—opposed New York's Responsible AI Safety and Education (RAISE) Act. They argued the legislation could conflict with federal policy and impose overly broad restrictions on AI development [Gothamist, 2025]. Both groups urged Governor Hochul to veto the bill, warning it could hamper innovation and create regulatory fragmentation [CCIA, 2025].	**California's 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 [DocumentCloud, 2024; TechCrunch, 2024]. Alongside other tech firms, Meta urged lawmakers to adopt more flexible, federally aligned policies [Carnegie Endowment, 2024]. **Responsible AI Safety and Education (RAISE) Act** In 2025, Meta opposed New York's Responsible AI Safety and Education (RAISE) Act through multiple affiliated groups. Tech:NYC, a trade group co-founded by Meta, warned the bill could restrict innovation and conflict with federal policy [Gothamist, 2025]. The AI Alliance also sent a letter to state leaders opposing the bill's scope and regulatory approach [AI Alliance, 2025]. The Computer & Communications Industry Association (CCIA), whose members include Meta, urged Governor Hochul to veto the legislation [CCIA, 2025].	**California's 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 [DocumentCloud, 2024; Carnegie Endowment, 2024]. The company also argued that the bill could raise national security risks by driving advanced research abroad [The Verge, 2024; Financial Times, 2024].	**California's SB 1047** In 2024, xAI CEO Elon Musk publicly supported the bill in an X post, stating: "This is a tough call and will make some people upset, but, all things considered, I think California should probably pass the SB 1047 AI safety bill. For over 20 years, I have been an advocate for AI regulation, just as we regulate any product/ technology that is a potential risk to the public." [Tech Crunch, 2024].	No publicly available information found

Table continues on next page

Sub-Indicator	Anthropic	DeepSeek	Google	Meta	OpenAI	x.AI	Zhipu AI
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 [New York Times, 2025]. The company also lobbied against the Trump-backed "Big Beautiful Bill," which aimed to override state AI regulation [WinBuzzer, 2025; Semafor, 2025].	No publicly available information found	In 2025, Google DeepMind supported federal preemption of state AI laws, urging a unified national framework to avoid regulatory fragmentation. In its response to the U.S. AI Action Plan, it called for federal leadership over issues like copyright, export controls, and development standards, warning that state-level rules could hinder innovation [Google Policy Response, 2025; TechCrunch, 2025].	In 2025, Meta supported federal preemption of state-level AI regulations, warning that fragmented laws could create compliance challenges and hinder innovation across jurisdictions [Meta, 2025]. The company's position aligned with broader industry efforts to shift AI governance to the federal level, drawing criticism from digital rights groups who argued this would weaken stronger state protections [X, 2025].	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 [OpenAI, 2025]. The company expressed concern that inconsistent state regulations could impose conflicting requirements and slow progress in the field [Bloomberg Law, 2025; Masood, 2025].	No publicly available information found	No publicly available information found

Indicator

Company Structure & Mandate

Definition & Scope

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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Uncommon governance structure. Fine-tuned 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]. This is meant to ensure board decisions can prioritize long-term safety and public benefit over short-term commercial pressures when making high-stakes decisions about transformative AI. 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].	For-profit company	Part of Google, a public for-profit company	Public for-profit company	Uncommon governance structure. Founded as a 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 is legally bound to pursue the Nonprofit's mission. 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." The for-profit arm has a 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 an '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 criticism. 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 that 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]	Filed as a Nevada for-profit benefit corporation. Definition by Secretary of State: "for-profit entities that consider the society and environment in addition to fiduciary goals in their decision-making process, differing from traditional corporations in their purpose, accountability, and transparency.". Registered purpose: "to advance human scientific discovery and deepen understanding of the universe." Nevada gives the state attorney-general independent standing to sue a public-benefit corporation that drifts from its mission, while Delaware does not [The Information; The Review Stories, 2025; NVSOS, 2014].	For-profit company

Sources

- Hendrycks, Dan. Introduction to AI safety, ethics, and society. Taylor & Francis, 2025. Section 8.4: Corporate Governance

Whistleblowing Protections

Indicator

Whistleblowing Policy Transparency

Definition & Scope

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 artefacts 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:

1. No transparency
2. Fragments public: Parts of the design of the whistleblowing policy are public
3. 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 behaviour 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 the firm's adherence to it. 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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Fragments Public Voluntary commitment in the safety framework [Anthropic, 2025], and comment on implementation status [Anthropic, 2024].	No transparency	**Fragments public** Employees are covered by Alphabet's group-wide code of conduct [Alphabet, 2024].	**Fragments public** WB policy referenced in Code of Conduct [Meta, 2024], integrity line available [Meta], and harassment policy is public in full [Meta].	**Full Policy Public** Details shared via FLI AI Safety Index Survey [Response] (16/16 relevant questions answered) Full Raising concerns policy public [OpenAI, 2024]; Integrity line available [OpenAI].	Details shared via FLI AI Safety Index Survey [Response] (16/16 relevant questions answered)	No transparency

Sources;

- Bullock, Charlie et al. "Protecting AI Whistleblowers." Lawafe. 2025. Accessed 1 July 2025.
- Wilson, Claudia et al. Whistleblower Protections for AI Employees. Center for AI Policy, 19 Jun. 2025,

Indicator

Whistleblowing Policy Quality Analysis

Definition & Scope

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, 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 line of defense when internal incentives 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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI	ISO 37002 "gold standard"
Overall average		**2.44**	**0**	**1.54**	**2.44**	**3.56**	**2.32**	**0**	**8.35**
1. Reporting Channels, Access, and Coverage		2.6	0	4.9	5.1	6.3	1.8	0	9.5
1.1 Protected Persons Coverage	Policy should at least cover current and former employees, contractors, shareholders, suppliers, former/prospective employees, and facilitators of reports	2	0	3	2	3	2	0	10
1.2 Policy Accessibility	Policy is easily accessible to all covered persons	0	0	2	8	8	0	0	10
1.3 External Reporting Information & Rights	Policy must provide clear information about external reporting channels and the right to approach these independently of internal processes, and explain or at least link to whistleblower protection rights	0	0	5	5	7	3	0	9
1.4 Multiple Reporting Channels	Offer multiple channels for reporting misconduct internally, incl. written, oral, and in-person	5	0	9	9	7	2	0	10
1.5 Anonymous Two-Way Reporting	The system enables fully anonymous reporting with secure two-way communication between the reporter and the investigators	4	0	5	9	10	0	0	9
1.6 Ombudsperson Channel	The reporting channel is operated by an outsourced whistleblowing service provider.	0	0	0	0	3	0	0	10
1.7 Executive Oversight Channel	A separate reporting channel is available for reports concerning senior executives (e.g., direct reporting line to the board audit committee) or board members	7	0	10	3	5	0	0	8
1.8 Broad but clear material scope	Material scope covers, at a minimum, potential violations of law and, code of conduct. Ideally, also further, broad categories, while retaining a high degree of clarity of what is in and out of scope.	3	0	5	5	7	7	0	10
2. Whistleblower Protections & Anti-Retaliation Measures		1.3	0	1.3	2.9	4	3.4	0	8.3
2.1 Confidentiality Protection	Strict protection is required for the reporter's identity and any third parties mentioned in reports	0	0	2	10	8	0	0	10
2.2 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	0	0	0	0	0	0	0	3
2.3 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	2	0	2	2	2	5	4	0
2.4 Post-Investigation Monitoring	Active monitoring for retaliation continues for a minimum of 12 months after the investigation concludes	0	0	0	0	0	0	0	8
2.5 NDA/Non-Disparagement Exceptions	Explicit statement that NDAs and non-disparagement agreements cannot prevent safety-related whistleblowing	7	0	0	0	7	10	0	10

Table continues on next page

Title	Description	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI	ISO 37002 "gold standard"
2.6 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	0	0	5	8	8	10	0	10
2.7 Handler/Investigator Protection	Explicit protections for employees who receive, investigate, or support whistleblowing reports	0	0	0	0	0	0	0	8
3. Investigation Process & Standards		0.8	0	1	3.2	2.2	0.4	0	7.6
3.1 Designated Impartial Receiver	A provably independent person or department must be designated to receive and handle reports, ideally attached to the board	4	0	5	6	6	2	0	9
3.2 Seven-Day Acknowledgment	Written confirmation of report receipt must be provided within 7 days	0	0	0	10	0	0	0	10
3.3 Three-Month Feedback Timeline	Investigation status and follow-up measures must be communicated to the reporter within 3 months	0	0	0	0	0	0	0	6
3.4 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	0	0	0	5	0	0	9
3.5 Investigation Appeal Process	Formal right to appeal investigation outcomes to an independent review body or board committee	0	0	0	0	0	0	0	4
4. System Governance & Quality Assurance		3	0	0	1	1	0	0	8
4.1 Comprehensive Effectiveness Metrics	Regular measurement tracking report outcomes, investigation timeliness, appeal rates, % of anonymous reports, retaliation incidents, and reporter satisfaction - not just volume	7	0	0	0	0	0	0	10
4.2 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	8
4.3 Secure Documentation System	Comprehensive audit trail with secure case management system and defined retention policies	0	0	0	5	5	0	0	9
4.4 Comprehensive Training Programs	Regular, role-specific training is provided for all employees, specialized training for managers and investigators, ideally measuring training effectiveness.	0	0	0	0	0	0	0	10
4.5 Independent System Certification	Regular third-party audit and certification of the whistleblowing system's effectiveness and compliance	8	0	0	0	0	0	0	3
5. AI Safety-Specific Provisions		4.5	0	0.5	0	4.3	6	0	N/A
5.1 AI Safety Commitment Protection	Explicit protection for reporting violations of frontier safety frameworks (e.g., RSP, Preparedness Frameworks), public AI safety commitments, and internal safety policies	8	0	0	0	5	5	0	
5.2 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)	0	0	0	0	2	2	0	

Table continues on next page

Title	Description	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI	ISO 37002 "gold standard"
5.3 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.	5	0	0	0	5	10	0	
5.4 Adequacy of 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	5	0	2	0	5	7	0	

Sources

- European Parliament and Council of the European Union. Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the Protection of Persons Who Report Breaches of Union Law. Official Journal of the European Union, 26 Nov. 2019
- International Organization for Standardization. ISO 37002:2021 - Whistleblowing Management Systems — Guidelines. ISO, 2021
- Nowers, Ida., and Terracol, Marie. "Monitoring Internal Whistleblowing Systems - A framework for collecting data and reporting on performance and impact". 2025

Indicator

Reporting Culture & Whistleblowing Track Record

Definition & Scope

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 organisation'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 behaviour 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.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Statement on non-disparagement agreements (June 2024): Cofounder Sam McCandlish announced that the firm has been using standard non-disparagement agreements in severance agreements, but now considers this practice to be in conflict with their mission and has started removing them. Stated that former employees who signed a non-disparagement agreement are free to state that fact, raise concerns about safety at Anthropic, and that Anthropic would not enforce non-disparagement agreements in such cases [LessWrong].	None	(Note: covers all of Google, not only DeepMind) **Satrajit Chatterjee (2022-ongoing):** Engineering manager fired in March 2022 after challenging a paper published by Google about AI chip design capabilities. A California state judge in July 2023 rejected Google's request to dismiss his wrongful termination and whistleblower protection claims [Bloomberg, 2023]. Chatterjee alleged that Google terminated him in retaliation for refusing to participate in what he viewed as misrepresentation of the company's AI technology capabilities, potentially defrauding shareholders and the public. Google stated the allegations were 'academic disputes' and defended the paper's scientific merit [The Star, 2023]. **Shareholder motion for stronger protections (2021):** Trillium Asset Management (Alphabet shareholder) filed a resolution calling for expanded whistleblower protections for Google employees. The resolution requested that Alphabet's Board of Directors oversee a third-party review of current whistleblower policies, citing the importance of strong protections for employees who raise concerns. *Outcome*: Alphabet's board recommended "AGAINST"; at the 2 June 2021 AGM, only **10% of total votes** (≈ 63.8 m) supported the motion [SEC, 2021]. **Margaret Mitchell (2021):** Co-lead of the ethical AI team, ran scripts to archive emails documenting the handling of Gebru's case. Fired for "exfiltrating thousands of files" in breach of security policy, according to Google [The Verge, 2021; MIT Technology Review, 2020].	(Note: covers all of Meta, not only Llama teams) **Sarah Wynn-Williams (2025):** Former global public policy director. Published a memoir and testified to Congress about Meta's alleged cooperation with China's government and misleading of lawmakers. Meta invoked a 2017 severance non-disparagement clause, won an emergency arbitration order potentially temporarily barring "disparaging" statements and blocking meetings with US/UK/EU legislators. Wynn-Williams testified before the Senate. Meta told TechCrunch that the arbitration order does not prohibit her from speaking to Congress and that the company does not intend to interfere with her legal rights [TechCrunch, 2025; CNN, 2025]. In Apr 2025, Sen. Grassley's letter to Meta demanding answers on NDAs allegedly silencing whistleblowers [Grassley, 2025]. **Internal memo threatens termination for leaks (Jan 31, 2025):** After CEO Mark Zuckerberg complained that "everything I say leaks," Meta CISO Guy Rosen circulated an internal memo warning that "we will take appropriate action, including termination," against employees who leak confidential information. The memo confirmed Meta had "recently terminated relationships with employees who leaked confidential company information." (The warning memo itself was subsequently leaked to the press.) [The Verge, 2025; Fortune, 2025]. **NLRB Ruling (2024):** The NLRB judge ruled that Meta's separation agreements used during the 2022 mass layoffs were illegal. The agreements affected over 7,000 employees who were required to sign "unlawfully overbroad" non-disparagement and confidentiality clauses in exchange for enhanced severance pay. This followed the precedent set by the McLaren Macomb decision in February 2023 [The Register, 2024; HRD, 2024]. **20 Employees terminated (2024):** ~20 employees terminated for leaks of internal meeting details; Meta said more firings may follow [TechCrunch, 2025].	**Dissolution of AGI readiness team (Oct 2024):** Team leader Miles Brundage left the firm. As part of a broader farewell message shared that some of his colleagues seem to think "that speaking up has big costs, and that only some people are able to do so.", but that he " think[s] people almost always assume that it's harder/more costly to raise concerns or ask questions than it actually is."[X, 2024]. The AGI readiness team was then dissolved within OpenAI [CNBC, 2024]. Brundage's exit then spurred former team member Rosie Campbell to depart [The Byte, 2024]. **Anonymous whistleblowers (Jul 2024):** Letter & formal SEC complaint allege OpenAI's NDAs illegally bar staff from alerting regulators and waive Dodd-Frank rewards. SEC matter pending [Tech Crunch, 2024; The Hill, 2024]. **Right-to-Warn" open letter – 11 current & former staff, plus peers at other firms (Jun 2024):** Called for an enforceable right to disclose AI-risk concerns without retaliation or broad NDAs. All current employees chose to remain anonymous. The letter cites criticism of OpenAI's equity claw-back clause [Right to Warn, 2024]. **Jan Leike (May 2024):** Superalignment Co-lead. Resigned, stating "safety culture and processes have taken a backseat to shiny products," and that the team lacked compute [X, 2024]. Co-team lead Sutskever left simultaneously to start competitor firm 'Safe Superintelligence.' The superalignment team was then disbanded [X, 2024]. Policy researcher Gretchen Krueger announces her departure hours later, stating similar concerns [The Verge, 2024]. **Exit-agreement overhaul after media leak (May 2024):** Vox revealed strict severance papers that let OpenAI *cancel vested equity* if ex-staff "disparaged" the company. After a media exposé, OpenAI **removed the clauses** and said it had never clawed back equity; CEO Sam Altman apologized, though leaked paperwork later showed his prior sign-off on the wording [Vox, 2024; The Verge, 2024]. Safety researcher Todor Markov left the company over the issue, arguing the debacle incident proved Altman was a person of low integrity who had directly lied to employees" [Futurism, 2025]. **Leopold Aschenbrenner (Apr 2024):** Researcher on Superalignment team says he was fired for circulating a memo to board members about security gaps; OpenAI says it was for leaking confidential info [Business Insider, 2024].	None	None

Table continues on next page

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
		Timnit Gebru (2020): Co-lead of Google's ethical AI team. Objected to Google's demand to retract a paper outlining large-language-model risks (bias, emissions). Google says she "resigned"; Gebru says she was **terminated**. Incident provoked >2,000-employee petition [ABC News, 2020; MIT Technology Review, 2020; Time, 2022]. **Mustafa Suleyman (2019, precedent for accountability):** Internal probe found patterns of workplace bullying. **Placed on administrative leave** July 2019; later moved to Google product role and eventually left Alphabet in 2022 [TechBrew, 2021].	**Arturo Bejar (2023):** Former engineering director. Testified before Congress that leadership, including. Mark Zuckerberg ignored evidence that Instagram harms teens (bullying, self-harm). He had emailed Zuckerberg, Sandberg, and other executives in 2021 with research showing harmful effects on young users. Meta stated it "does not agree" with Bejar's characterisation and highlighted existing safety tools; no legal action has followed [CNBC, 2023; NPR, 2023; OPB, 2023]. **Frances Haugen (2021):** Former product manager. Supplied thousands of internal files ("Facebook Papers") to the SEC & US Congress, alleging Meta misled the public about known harms (teen mental health, misinformation). Meta said documents were "cherry-picked" and Haugen's claims "mischaracterise" its work. No litigation between parties. Haugen testified before the Senate Oct 2021 [CBS, 2021] **Sophie Zhang (2020-21):** Data scientist. Farewell memo described government-backed political manipulation campaigns across 25 countries on Facebook; reposted the memo on a password-protected personal site. Facebook deleted her internal post and requested her web-host & registrar remove the external site, which they did. Zhang declined a severance agreement containing a non-disparagement clause and later testified before the British Parliament and provided documents to US law enforcement [Independent, 2021; BuzzFeed, 2020].	**Daniel Kokotajlo (Apr 2024):** Governance researcher resigned because he "Lost confidence [OpenAI] would behave responsibly around AGI" [Futurism, 2024]. **William Saunders (Feb 2024):** Interpretability engineer resigned, telling *Business Insider* leadership "was not adequately addressing" catastrophic-risk issues; later co-signed the "Right-to-Warn" letter [Business Insider, 2024]. **Board coup & reversal (Nov 2023):** Board unexpectedly removed CEO, stating he "was not consistently candid in his communications". Altman was reinstated within a week. A WilmerHale special review later found his conduct "did not mandate removal" [ARS Technica, 2023]. In the aftermath, three independent directors (including Helen Toner) resigned [Aljazeera, 2024]. Toner later stated, "For years, Sam had made it really difficult for the board to actually do that job by, you know, withholding information, misrepresenting things that were happening at the company, in some cases outright lying to the board.[..]". [TED, 2024].		

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Exemplary accountability ensures safety-focused decision-making at all levels
- **B** Strong accountability enables safety-focused decision-making
- **C** Moderate accountability with gaps affecting safety decision-making
- **D** Weak accountability hinders safety-focused decision-making
- **F** No meaningful accountability

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

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.

Table of Contents

Technical Specifications

Indicator

System Prompt Transparency

Definition & Scope

This indicator reports whether companies publicly disclose the system prompts used in their most capable deployed AI models. Evidence includes published system prompts in model cards, technical documentation, or dedicated transparency pages, and changelogs. Best practice involves publishing exact prompts used in production, version history, verification that prompts are used in production, and explanations of design choices.

Why This Matters

System prompts fundamentally shape AI behavior and safety properties, yet most companies keep them secret. Publishing prompts enables researchers to verify to better understand the models and makes the company's intended behaviour transparent. High transparency can reflect a commitment to accountability.

Anthropic	**Transparency** In March 2024, Anthropic shared the full system prompt alongside the release of Claude 3 as a one-off [Fast Company, 2024]. Since August 2024, Anthropic has publicly shared the systems' prompts for the Claude.ai web interface and mobile apps since August 2024. Shared system prompts for six models, plus several updates. They further committed to logging changes they make they make to these prompts online. Shared systems prompts do NOT currently cover the API [TechCrunch, 2024; X, 2024; Anthropic]. Simon Willison reported that the publicly shared version does not include the description of various tools available to the model [Simon Willison, 2025].
DeepSeek	Frontier model weights are public, so the system prompt can be decided by user/hosting service. Their own hosted service does not disclose it.
Google DeepMind	No transparency on system prompts for Frontier Systems.
Meta	Frontier model weights are public, so the system prompt can be decided by the user/hosting service. Their own hosted service does not disclose it.
OpenAI	No transparency on system prompts for Frontier Systems.
x.AI	**Transparency:** Following the incident in May 2025 listed below, x.AI published their system prompts for Grok (on xAI & X) on Github and promised these will be regularly updated [Github, 2025]. **Incidents:** February 2024: A change to the system prompt, Grok briefly censored responses about Elon Musk and Donald Trump spreading disinformation. After the issue received public attention, xAI quickly reverted the changes and publicly stated that the problem was caused by an unnamed employee conducting unauthorized modifications [Fortune, 2024]. May 2025: After a change to the system prompt, Grok started randomly discussing whether there was a "white genocide" happening in South Africa in many completely unrelated conversations. The AI Chatbot told users it was 'instructed by my creators' to accept 'white genocide as real and racially motivated' [Guardian, 2025]. x.AI quickly apologized for this incident and rolled back the changes. They reported that unauthorized modifications by an employee caused the incident [X, 2025].
Zhipu AI	Frontier model weights are public, so the system prompt can be decided by the user/hosting service. Their own hosted service does not disclose it.

Sources

- Kokotajlo, Daniel, and Alexander, Scott. "Make The Prompt Public." AI Futures Project, 17 May 2025

Indicator

Behavior Specification Transparency

Definition & Scope

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

Model specifications reveal how companies intend their AI systems to behave in complex situations, including safety-critical decisions. 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.

Anthropic	**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's it 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 words) 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." **Benefits:** Readable, transparent, and explicitly formulated principles, as opposed to RLHF, which leverages implicit values. **Limitations:** Version uncertainty: Only the May 2023 constitution is public; the current production versions are unknown Anthropic uses a "variety of techniques including human feedback, Constitutional AI [..], and the training of selected character traits.". Given that other approaches are incorporated in post-training, the impact of any one of them is unclear. 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. **Source:** [Anthropic, 2025]
DeepSeek	No detailed specification available, but frontier model weights are public, so models can be modified.
Google DeepMind	No detailed specification available
Meta	No detailed specification available, but frontier model weights are public, so models can be modified.
OpenAI	**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's it for:** 1) Human RLHF guidance – provides a single, public rule-book that labelers follow when creating preference data.

Table continues on next page

	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 principal 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 potential, 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 **Key Benefits** Greater transparency of intended model behavior. Finer-grained steerability via the chain of command Reduced reliance on implicit human values; models can show interpretable reasoning steps grounded in the Spec. **Transparency & Limitations** Production models don't fully reflect the spec yet. OpenAI states: "While the public version of the Model Spec may not include every detail, it is fully consistent with our intended model behavior." **Source:** [OpenAI, 2025]
x.AI	No detailed specification available
Zhipu AI	No detailed specification available, but frontier model weights are public, so models can be modified.

Sources

- Kokotajlo, Daniel, and Alexander, Scott. "Make The Prompt Public." AI Futures Project, 17 May 2025
- Ball, Dean. "4 Ways to Advance Transparency in Frontier AI Development." The Foundation for American Innovation, 16 Oct. 2024

Voluntary Cooperation

Indicator

G7 Hiroshima AI Process Reporting

Definition & Scope

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	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Substantive submission [OECD AI, 2025]	(Not based in a G7 nation)	Substantive submission (Google) [OECD AI, 2025]	No Submission	Substantive submission [OECD AI, 2025]	No Submission	(Not based in a G7 nation)

Sources

- Perset, Karine, James Gealy, and Sara Fialho Esposito. "Shaping Trustworthy AI: Early Insights from the Hiroshima AI Process Reporting Framework." OECD.AI, 11 Jun. 2025
- Ministry of Internal Affairs and Communications, Japan. G7 Hiroshima Process on Generative Artificial Intelligence. Ministry of Internal Affairs and Communications, Japan
- Organisation for Economic Co-operation and Development (OECD). OECD.AI Policy Observatory: Reports. OECD

Indicator

FLI AI Safety Index Survey Engagement

Definition & Scope

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

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
None	None	None	None	Survey response submitted [PDF]	Survey response submitted [PDF]	Survey response submitted [PDF]

Risks & Incidents

Indicator

Serious Incident Reporting & Government Notifications

Definition & Scope

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.

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	xAI	Zhipu AI
Serious incident reporting frameworks						
	Chinese AI firms operate under several regulations with mandatory incident reporting requirements, often under short timeframes. We list applicable GenAI specific frameworks but not those focused on (data-/cyber-) security: - Interim Measures for Generative AI Services, Art. 14 (Aug 2023) – Gen-AI providers that detect illegal content or model misuse must "promptly" stop generation, rectify, and inform the competent authorities [China Law Translate, 2023] - Deep-Synthesis Provisions (Jan 2023) – Deep-fake service providers must remove illegal or harmful synthetic content, preserve records, and "timely" report the incident to the CAC and other competent departments [Cyberspace Administration of China, 2023]					Chinese AI firms operate under several regulations with mandatory incident reporting requirements, often under short timeframes. We list applicable GenAI specific frameworks but not those focused on (data-/cyber-) security: - Interim Measures for Generative AI Services, Art. 14 (Aug 2023) – Gen-AI providers that detect illegal content or model misuse must "promptly" stop generation, rectify, and inform the competent authorities [China Law Translate, 2023] - Deep-Synthesis Provisions (Jan 2023) – Deep-fake service providers must remove illegal or harmful synthetic content, preserve records, and "timely" report the incident to the CAC and other competent departments [Cyberspace Administration of China, 2023]
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" [Anthropic, 2025].		Frontier Safety Framework 2.0 states that if a model reaches a "Critical Capability Level" posing unmitigated material risk, DeepMind "aims to share information with appropriate government authorities" and may also notify other external organisations [Google, 2025].				

Table continues on next page

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	xAI	Zhipu AI
Public transparency reports						
Anthropic published one comprehensive misuse report, which documents real-world cases of actors attempting to exploit Claude for malicious purposes, along with detection methods and enforcement actions taken. -Mar 2025 – "Misuse Monitoring and Response Report" [Anthropic, 2025]. -Platform Security transparency page provides some enforcement statistics, including banned accounts for Usage Policy violations, number of appeals processed, CSAM reports to NCMEC, and law enforcement requests [Anthropic, 2024].		Published a detailed report on 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 -Jan 2025 - 'Adversarial Misuse of Generative AI" [Google 2025].	Meta consistently issues quarterly integrity reports about its **platforms** [Meta, 2024], which include reports on disrupting adversarial threats such as influence operations [Meta, 2025]. No reports for frontier AI models are available.	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" [OpenAI, 2024] - May 2024 – "Disrupting a Covert Iranian Influence Operation" [OpenAI, 2024] - Jun 2024 – "Update on Disrupting Deceptive Uses of AI" [OpenAI, 2024] - Aug, 2024: "Disrupting a covert Iranian influence operation" [OpenAI, 2024] - Oct 2024 – "Influence and cyber operations: an update" [OpenAI, 2024] - Feb 2025 - "Disrupting malicious uses of our models" [OpenAI, 2025] - Jun 2025 - Disrupting malicious uses of AI [OpenAI, 2025]		

Table continues on next page

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	xAI	Zhipu 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 the 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].		The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate the 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].	The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate the 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].	The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate the 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].		Zhipu AI is a founding member of the IIFAA "Trusted Agent Inter-connect Working Group" (Dec 2024) alongside Huawei, Alibaba, ByteDance, etc.; the group sets cross-agent security and data-sharing norms [China Daily, 2024].

Table continues on next page

Indicator

Extreme-Risk Transparency & Engagement

Definition & Scope

Assesses the extent to which companies and their leadership (A) publicly recognise the potential for catastrophic AI harm and (B) proactively disseminate evidence-based analyses of such risks to external stakeholders. 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).

Note: The research methodology for this indicator did not follow a formal structure; results are incomplete and likely biased/skewed by the prominence of different media reports.

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	The company and its leaders regularly and proactively communicate extreme risks. Examples from CEO Dario Amodei: - 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 AI Risk signed by: Dario Amodei (CEO), Daniela Amodei (President), Jared Kaplan (co-founder), Chris Olah (co-founder) Relevant blogs by Anthropic are below. Several share quantitative evidence related to extreme risks: ▪ Progress from our Frontier Red Team [Anthropic, 2025] ▪ Third-party testing as a key ingredient of AI policy [Anthropic, 2024] ▪ Reflections on responsible scaling policy [Anthropic, 2024] ▪ The case for targeted regulation [Anthropic, 2024] ▪ Frontier Threats Red Teaming for AI Safety [Anthropic, 2023]
DeepSeek	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].

Table continues on next page

Google DeepMind	Corporate communications rarely mention extreme risks. Google DeepMind's leadership regularly discusses extreme risks in media interviews. Google's leadership does not. **Leadership examples:** ▪"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]. Time reported Hassabis saying: "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; Dwarkesh Podcast, 2024; TIME, 2023]. Shane Legg (Chief AGI Scientist) communicates a similar stance [Dwarkesch Podcast; 2023, Google DeepMind, 2023]. Talking to Axios, Legg recently stated AI is a very powerful technology, and it can and should be regulated." [Axios,2025]. Google's CEO, Sundar Pichai, stated that "The biggest risk could be missing out," at the AI Action Summit in Paris yesterday https://observer.com/2025/02/biggest-risk-ai-is-missing-out-google-ceo-sundar-pichai/?utm_source=chatgpt.com CAIS statement on AI Risk signed by: Demis Hassabis (CEO), Shane Legg (Co-Founder), Lila Ibrahim (COO)
Meta	Company and leadership rarely address extreme risks. Mark Zuckerberg and Chief AI Scientist Yann LeCun express the strongest counternarrative 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].
OpenAI	OpenAI and its leadership sometimes talk about extreme risks CEO Altman's communications have changed over time. 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 notably 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]. His recent blogs focus on the benefits Superintelligence could bring [Altman, 2025]. CAIS statement on AI Risk signed by: Sam Altman (CEO), Adam D'Angelo (board member), Wojciech Zaremba (cofounder) Relevant blogs: ▪ Preparing for future AI capabilities in biology: Acknowledges AI could enable people to: "recreate biological threats or assist highly skilled actors in creating bioweapons.", then explains OpenAI's approach to preventing misuse [OpenAI, 2025].

Table continues on next page

x.AI	xAI itself does not publicly share information about extreme risks. CEO Musk has a track record of raising concerns. 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 (cofounder), Tony Wu (co-founder). Musk signed the FLI pause letter [FLI 2023]
Zhipu AI	Corporate communications don't speak about the potential for extreme risks. Leadership is discussed publicly. Tang Jie 唐杰 (Chief Scientist) signed a 2024 track 2 diplomacy statement acknowledging potential for catastrophic risks: "Collectively, we must prepare to avert the attendant catastrophic risks that could arrive at any time." [IDAIS, 2024] When speaking about AGI at the Seoul Summit, CEO Peng said: "[..] crucial responsibility of ensuring AI safety. As we delve deeper into the realms of AGI, it is imperative that we prioritize the development of robust safety measures to align AI systems with human values and ethical standards, thereby safeguarding our future in an AI-driven world.' [UK Gov, 2024] Zhang Peng (CEO) was the only industry representative among Chinese scientists signing the IDAIS statement on AI safety redlines that should not be crossed, which stated: "[..] AI systems may pose catastrophic or even existential risks to humanity within our lifetimes." [IDAIS, 2024; Carnegie Endowment, 2024]. He gave a speech emphasizing the need for research to align superintelligent systems [36kr, 2024].

TO BE COMPLETED BY PANELLISTS

Grading

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

	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Grades	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**	**Grade**
Grade comments (Justifications, opportunities for improvements, etc.)							

Grading Scales

Grading scales are provided to support consistency between reviewers.

- **A** Exemplary transparency enables informed safety decisions by all stakeholders
- **B** Strong transparency supports effective oversight and public understanding
- **C** Moderate transparency with selective disclosure patterns
- **D** Limited transparency hinders risk assessment
- **F** Deceptive or negligible information sharing

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 (15 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 the protections available to current full-time employees. You may explain variations for different stakeholder groups in the final question.

You can use the textbox 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 are 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	Zhipu AI	xAI	OpenAI
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:	Prefer not to answer (skips whistleblowing section)	▪ Non-public policy exists - Please briefly explain your rationale for keeping it private: ▪ Only applies to xAI employees	Public WB policy - Please provide URL here: https://openai.com/index/openai-raising-concerns-policy/ https://cdn.openai.com/policies/raising-concerns-policy-blog-copy-202410.pdf
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):		HR department	Board/Audit Committee, Compliance/Legal department, HR department HR, board/audit as well
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 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, 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,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

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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
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	Current employees,Contractors and self-employed workers
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):		Human Resources department,Direct disclosure to a statutory or supervisory authority	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
For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)	▪ Limited Reporting Channels (Former employees | Contractors) ▪ Limited Reportable Issues (Former employees | Contractors) ▪ Limited Retaliation Protection (Former employees | Contractors) ▪ No Limitations (Former employees | Contractors)	▪ Limited Reporting Channels () ▪ Limited Reportable Issues () ▪ Limited Retaliation Protection () ▪ No Limitations ()	▪ Limited Reporting Channels (Former employees | Contractors) ▪ Limited Reportable Issues () ▪ Limited Retaliation Protection (Former employees | Contractors) ▪ No Limitations ()	▪ Limited Reporting Channels (Contractors: Some channels, such as speaking to your current HR representative, are inherently available only to current employees.) ▪ Limited Reportable Issues () ▪ Limited Retaliation Protection () ▪ No Limitations ()
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 does not provide for anonymous reporting	Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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):		▪ Imminent risk of serious harm,Other (specify): ▪ Reasonable fear of physical harm	
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):		None of the above	Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary
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

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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		Termination/Dismissal,Demotion, or negative performance reviews,Reduction in compensation or benefits,Blacklisting within the industry,Legal action against the whistleblower	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 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.
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		No - we do not include such restrictions in our agreements	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.

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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		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)

External Pre-Deployment Safety Testing (6 Questions)

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

Frontier models:

- Anthropic - Claude 4 Opus
- DeepSeek - R1
- Google DeepMind - Gemini 2.5 Pro
- Meta - Llama 4 Maverick
- OpenAI - o3
- xAI - Grok3
- Zhipu AI - GLM-4 Plus

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

Question Title	Available options	Zhipu AI	xAI	OpenAI
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. 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): We intend to share our model with certain independent organizations for evaluation purposes; however, we prefer not to disclose their identities.	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): We've worked with the US and UK AI Safety Institutes, and independent third party labs such as METR, Apollo Research, and Pattern 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 o3 System Card. Third party assessors were provided OpenAI o3 early checkpoints, as well as the final launch candidate models to conduct their assessments. As part of our ongoing efforts to consult with external experts, OpenAI granted early access to these versions of o3 to the U.S. AI Safety Institute to conduct evaluations of the models' cyber and biological capabilities, and to the U.K. AI Security Institute to conduct evaluations of cyber, chemical and biological, and autonomy capabilities, and an early version of the safeguards. METR measured the models' general autonomous capabilities, and reward hacking. Pattern Labs evaluated the model's cybersecurity related capabilities (evasion, network attack simulation, and vulnerability exploitation). Apollo Research evaluated in-context scheming and strategic deception. In some instances we paid private consultants for their work, but payment is not conditioned on the content of their findings.
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	Inference API with safety filters disabled (no inference-time mitigations)		▪ 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)
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	>3 weeks		>2 weeks

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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.,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.		▪ 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 red teaming period are under NDA / require prior approval ▪ 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 red teaming period ▪ 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.
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.	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 rates can depend on technical feasibility in some cases.
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): We will communicate with the evaluators to confirm whether it is permissible to retain relevant records.		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).

Table continues on next page

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	Zhipu AI	xAI	OpenAI
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):	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.
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):	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.	No formal risk management requirements for internal deployments	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.
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):	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,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	Logging all inputs and outputs from internal use and retaining them for at least 30 days	See answer to Q24, above.

Safety Practices, Frameworks, and Teams (9 Questions)

Question Title	Available options	Zhipu AI	xAI	OpenAI
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 – we released the same model version.	Yes – we released the same model version. Yes. We ran our evaluations on an earlier checkpoint and then confirmed our automated evaluation results on the final checkpoint.
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.*	This matter is considered company confidential, and we prefer not to answer.	Team name: AI Safety Engineer Mission and scope: Forward-looking / fundamental research + model improvements such as jailbreak prevention and safety classifiers FTEs: Three Team name: Product Safety Mission and scope: Immediate product safety such as jailbreak prevention FTEs: One	We have multiple teams across safety research focused on safety, alignment, evaluations, trustworthiness and governance.
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. ▪ Other (please briefly describe):	2) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority.	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.

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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	Published & Implementation in progress	Published & Implementation in progress
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)	Yes, internally. (Please briefly explain why you have not published it) Currently, Zhipu's models have not yet reached the level of AGI, so we prefer not to release the related plans.	No, but we're working on it	Yes, internally. (Please briefly explain why you have not published it) For more on our approach to ensuring that AGI remains controllable and safe, see https://openai.com/safety/how-we-think-about-safety-alignment/

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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	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.,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.	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.	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, 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. 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 and continues to iterate on 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.

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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. ▪ 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	**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.	**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.'	**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 publicly urged against optimizing on chains of thought: https://openai.com/index/chain-of-thought-monitoring/
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).	Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below).	None – no TSFT safety testing performed (skips follow-up).	None – no TSFT safety testing performed (skips follow-up). None. 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.
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):	Other (please specify): Biological, Persuasion, Chemical, Cyber-offense, Political		

Table continues on next page

Question Title	Available options	Zhipu AI	xAI	OpenAI
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: Securing Research Infrastructure for Advanced AI. ▪ Third party collaboration on security: OpenAI maintains a bug bounty program through BugCrowd (https://bugcrowd.com/openai), and welcomes responsible disclosures from third parties via our coordinated vulnerability disclosure policy (https://openai.com/policies/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.

17th July 2025