Future of Life Institute | July 2026

AI Safety Index Summer 2026

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

Full report at: **futureoflife.org/index** | Contact us: **policy@futureoflife.org**

Full report

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Overall Grade	C+	C	C	D+	D-	D-	F	F	F
Score	2.66	2.28	2.01	1.32	0.88	0.87	0.65	0.47	0.33
Grade Trend (Winter 25)	C+	C+ ▼	C	D ▲	D ▼	D-	D ▼	D ▼	N/A
Risk Assessment (6 indicators)	C+	C+	C+	D+	F	F	D-	F	F
Current Harms (9 indicators)	B-	C	C	D-	C-	C-	F	D-	F
Safety Frameworks (4 indicators)	B-	C+	C	C-	D-	D-	D	F	F
Existential Safety (4 indicators)	D+	D+	D	F	F	F	F	F	F
Governance & Accountability (4 indicators)	B	C	C-	D+	F	D-	F	F	F
Information Sharing (10 indicators)	B+	B-	B-	D+	D	D	D	D-	D-
Survey Responses	✓	✓	✓	✓	✓	✗	✗	✗	✗

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

Executive Summary

- **Anthropic, OpenAI, and Google DeepMind stay on top.** Anthropic again earns the highest overall grade and leads five of six domains via relatively strong transparency, a comparatively established safety framework, technical research, and governance. OpenAI now leads in Risk Assessment on the strength of a broader evaluation suite and diverse engagement with external testing.
- **Meta improves and xAI deteriorates:** Meta improved from 6th to 4th place, while xAI dropped from 4th to 7th place.
- **Inadequate safety is a global problem, not a regional one.** Three companies receive failing grades, one each from the US (xAI), China (DeepSeek), and Europe (Mistral).

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.

- **Reviewers flagged the industry's pivot to military AI use as an emerging current harm risk.** From 2024 to 2026, companies including Anthropic, OpenAI, Google DeepMind, and Meta that previously banned military applications gradually reversed course, joining xAI and Mistral in actively seeking defense partnerships. Despite their limits on domestic surveillance and autonomous weapons, Anthropic drew criticism from the review panel for "questionable military engagements," including a reported link to the Minab school strike that caused mass civilian deaths. Leading Chinese firms, meanwhile, face U.S. allegations of military ties that Alibaba Cloud and Z.ai deny.
- **Even industry leaders in safety practices are retreating from prior commitments, despite calling publicly for a pause.** Anthropic, OpenAI, Google DeepMind, and Meta have weakened or voided pledges to pause unilaterally if certain red lines were approached. Reviewers call this "moving goalposts" and argue that it has "undermined safety frameworks across the board".
- **Existential Safety is the weakest domain industry-wide.** No company exceeds C-; most score D or below. Constructive attempts exist, such as Anthropic's constitutional classifiers, OpenAI's call for governance institutions, Google DeepMind's monitoring commitments, and Meta's loss-of-control provisions, but are judged by panelists to be "entirely inadequate." Dominant paradigms such as interpretability and Chain-of-Thought (CoT) monitorability are questioned because "detection is not prevention."

What is the AI Safety Index?

Leading AI systems are advancing rapidly, raising increasingly urgent questions about current harms and long-term controllability as models grow more autonomous, capable, and potentially self-improving. As capabilities grow, both the opportunities offered by these systems and the risks they pose expand accordingly. The AI Safety Index, developed by the Future of Life Institute with an independent panel of technical and governance experts, provides an impartial and biannual evaluation of how responsibly leading AI companies are approaching these challenges. Competitive pressures may reward profits over safety, so the Index aims to counterbalance those incentives by making companies' safety practices visible and comparable, creating reputational pressure to meet higher standards.

Methodology

The 2026 Summer AI Safety Index assesses the safety practices of nine AI companies – Anthropic, Alibaba Cloud, DeepSeek, Google DeepMind, Meta, Mistral, OpenAI, xAI, and Z.ai – across six critical domains, to foster transparency, promote robust safety practices, highlight areas for improvement and empower policymakers and the public to discern genuine safety measures from empty commitments.

An independent review panel of seven leading experts on technical and governance aspects of general-purpose AI volunteered to assess the companies' performances across 37 indicators of responsible conduct, contributing letter grades, brief justifications, and recommendations for improvement. The evaluation was supported by a comprehensive evidence base with company-specific information sourced from 1) publicly available material, including related research papers, policy documents, news articles, and industry reports, and 2) a tailored industry survey which firms could use to increase transparency around safety-related practices, processes and structures. The full list of indicators and collected evidence is presented in the full report.

Independent Review Panel

David Krueger: Assistant Professor in Robust, Reasoning, and Responsible AI, University of Montreal; Core Academic Member, Mila; Founder, Evitable

Sharon Li: Associate Professor of Computer Science, University of Wisconsin-Madison

Tegan Maharaj: Assistant Professor in Machine Learning, HEC Montréal; Core Academic Member, Mila

Sneha Revanur: Founder and President, Encode

Stuart Russell, OBE: Professor of Computer Science, UC Berkeley; Director, Center for Human-Compatible AI

Robert Trager: Director, Oxford Martin AI Governance Initiative

Yi Zeng: Professor, Gaoling School of AI, Renmin University of China; Founding Dean, Beijing Institute of AI Safety and Governance