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].				
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]		
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].