	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
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
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]