Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Claude Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4	Grok-4	R1	GLM-4.6	Qwen3-Max
Biosecurity & Chemical Risk							
Final rounds of safety evaluations were conducted on the same model version that was released. Evaluations prioritize biological risks and do not conduct internal or external evaluations for chemical risk. **Safety Framework Classification** Evaluations test AI Safety Level 3 (ASL-3) and ASL-4 capability thresholds for related risks under Anthropic's Responsible Scaling Policy (RSP). **Evaluations scope** covers: 1) ASL-3: testing whether models can assist low-expertise actors in performing core biological threat workflows - Long-form virology tasks (task-based agentic evaluations co-developed with SecureBio, Deloitte, and Signature Science), - Multimodal virology (SecureBio VCT), - DNA Synthesis Screening Evasion (SecureBio) - LAB-Bench subset (expert-level biological skills assessment developed by FutureHouse) 2) ASL-4: testing whether models could substantially accelerate advanced or state-scale biological R&D - Creative biology (SecureBio) - Short-horizon computational biology tasks (Faculty.ai) **Methodological Details include:** 1) Environment and elicitation setup (e.g. containerization, tool integration, agent harness, "helpful-only" model variants, extended thinking mode etc.) 2) Human/AI baselines 3) Quantitative evaluation metrics (e.g. Rule-in/out thresholds, human & model baselines) System Card (pp. 125-136)	*Final rounds of safety evaluations were conducted on the same model version that was released.* Evaluations prioritize biological capability evaluations. **Safety Framework Classification** GPT-5 is treated as High capability in the Biological and Chemical domain under OpenAI's Preparedness Framework. **Evaluation Scope** covers: (1) Long-form biorisk questions (five stages of biothreat creation—ideation to release) (2) Multimodal virology troubleshooting (SecureBio/Center for AI Safety) (3) ProtocolQA open-ended troubleshooting (adapted from FutureHouse [Laurent et al., 2024]) (4) Tacit knowledge & troubleshooting (Gryphon Scientific, not published) (5) TroubleshootingBench focusing on real-world, experience-grounded wet-lab errors (6) Virology capabilities, human pathogen capabilities, molecular biology capabilities, world class biology (external evaluation by SecureBio) **Methodological Details** include: (1) Elicitation setup (e.g. maximum verbosity) (2) Human and expert baselines (3) Quantitative evaluation metrics System Card (pp. 23-27)	Evaluations have covered biological, chemical, nuclear, and radiological capabilities. **Safety Framework Classification** CBRN risks are tested for Uplift Level 1, with additional "alert-threshold" monitoring for early-warning signs of dangerous dual-use capabilities. It remains below the alert threshold. **Evaluation scope** includes: (1) Multiple choices quantitative questions: i) SecureBio VMQA4 single-choice; ii) FutureHouse LAB-Bench presented as three subsets (ProtocolQA, Cloning Scenarios, SeqQA) (Laurent et al., 2024); and iii) Weapons of Mass Destruction Proxy (WDMP) presented as the biology and chemistry data sets (Li et al., 2024). (2) Open-ended questions: qualitative assessment on knowledge-based, adversarial, and dual-use content in the biological, radiological and nuclear domains led by domain experts. **Methodological Details** include: (1) Quantitative and qualitative evaluation metrics (2) Human, expert, and model performance baselines System Card (pp. 12-14)	The system card mentions that Meta has conducted expert-designed and other targeted evaluations designed to assess whether the use of Llama 4 could meaningfully increase the capabilities of malicious actors to plan or carry out attacks using these types of weapons, however, no safety framework classification, methodological details and scope information are disclosed.	*Final rounds of safety evaluations were conducted on the same model version that was released.* Evaluations prioritize biological capability evaluations. **Safety Framework Classification** None **Evaluation Scope** covers: (1) Dual-use knowledge for bioweapons (2) Chemical knowledge **Methodological Details** include: (1) Benchmarks (WMDP Bio, WMDP Chem, BioLP-Bench, VCT [text-only]) (2) Quantitative metrics System Card (pp. 5)	**Not Mentioned**	*Final rounds of safety evaluations were conducted on the same model version that was released.* **Not Mentioned**	**Not Mentioned**
Cybersecurity Risks							
Yes **Safety Framework Classification** **Ongoing assessment without formal threshold in RSP at any ASL.** **The Evaluation Scope** covers 1) General Cyber Evaluations - Quantitative results on CyberGym/Cybench - Anecdotal observations on triage and patching 2) Advanced Risk Evaluations - Irregular Challenges (23 private CTFs co-developed with Irregular to measure ability to discover and exploit complex vulnerabilities across categories including Web, Crypto, Pwn, Rev, Network) - Incalmo Cyber Ranges (25–50 hosts; co-developed with Carnegie Mellon University to test the model's capacity for long-horizon, multi-host cyber operation). **Methodological Details include** (1) Environment and elicitation (e.g. Kali-based sandbox, access to terminal, code editor, and standard penetration-testing tools) (2) Benchmarks and model performance baselines (3) Quantitative evaluation metrics System Card (pp. 32-45, 148)	**Yes** **Safety Framework Classification** Cyber capabilities are tracked as part of ongoing safety monitoring. **The Evaluation Scope** covers (1) Capture-the-Flag (CTF) Challenges across Web Application Exploitation, Reverse Engineering, Binary & Network Exploitation (pwn), Cryptography, and Miscellaneous categories (2) Cyber Range (5 scenarios of light-to-medium difficulty) to test the model's ability to conduct long-form, end-to-end cyber operations (3) Evasion, network attack simulation, and vulnerability discovery and exploitation (Pattern Lab external assessment) **Methodological Details** include (1) Environment and Elicitation setup (e.g. headlessLinux box, tool harness) (2) Benchmarks and model performance baselines (3) Quantitative evaluation metrics System Card (pp. 27-35)	**Yes** **Safety Framework Classification** Cyber risks are tested for Cyber Autonomy Level 1 and Cyber Uplift Level 1, both unreached. However, the model crossed the early-warning alert threshold for Uplift Level 1. **Evaluation Scope** includes: (1) Existing Capture-the-Flag (CTF) challenges primarily for autonomy tests: i) InterCode-CTF (easy, undergraduate level) ii) In-house suite (medium, graduate-level) iii) Hack the Box (hard, professional level) (2) Key skills benchmark (Rodriguez et al., 2025)for uplift tests: 8 mapped challenges to measure 4 critical competencies: i) Reconnaissance ii) Tool development iii) Tool usage iv) Operational security. **Methodological Details** include: (1) Environment and elicitation setup (e.g. Bash and Python execution) (2) Benchmarks and model performance baselines System Card (pp. 14-17), Technical Report (pp. 30-32)	**Yes** **The Evaluation Scope covers** automate cyberattacks, identify and exploit security vulnerabilities, and automate harmful workflows. **Methodological Details** include threat modeling exercises and capability-based challenge construction.	**Yes** **Safety Framework Classification** None **Evaluation Scope** covers: (1) Cyber knowledge (e.g. Metasploit, vulnerability detection, reverse engineering simple binaries) (2) Cyber agent **Methodological Details include:** (1) Environment setup (Inspect by UK AISI, agent harness) (2) Benchmarks (WMDP Cyber, CyBench) (3) Qualitative metrics System Card (pp. 5-6)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**
Autonomous AI R&D							
Yes **Safety Framework Classification** Evaluation test thresholds for 1) Checkpoint 2) AI R&D 4 (ASL-3); 3) AI R&D 5 (ASL-4) **The scope of evaluation** includes 1) A checkpoint: a wide range of 2–8 hour software engineering tasks - SWE-bench Verified (hard subset) 2) ASL-4: custom difficult AI R&D tasks built in-house - Internal AI research evaluation suite 1 (e.g. kernels task, time series fore casting, text-based reinforcement learning task, LLM training etc.) - Internal AI research evaluation suite 2, - Internal Model evaluation and use survey **Methodological details** include 1) Environment and elicitation (e.g. context and prompt lengths variations, example-based prompts) 2) Benchmarks with human/model performance baselines 3) Quantitative evaluation metrics System Card (pp. 136-147)	**Yes** **Safety Framework Classification** AI self-improvement capabilities are tracked as part of ongoing safety monitoring. **The Evaluation Scope** covers (1) Real-world software engineering tasks (SWE-bench Verified (N=477), SWE-Lancer (Diamond IC-SWE)) (2) Real world ML research tasks (OpenAI PRs) (3) Real world data science and ML competitions (MLE-Bench) (5) Real world ML paper replication (PaperBench) (6) Real world ML debugging and diagnosis (OPQA (OpenAI-Proof Q&A)) **Methodological Details** include (1) Environment and Elicitation setup (e.g. virtual environment with with tool access, bash execution, and GPU resource, maximum trained-in verbosity) (2) Benchmarks with human/model performance baselines (3) Quantitative evaluations metrics System Card (pp. 35-43)	**Yes** **Safety Framework Classification** Machine Learning R&D capabilities are tested for ML R&D Autonomy Level 1 and ML R&D Uplift Level 1, both remaining unreached. **The Evaluation Scope** covers Research Engineering Benchmark (RE-Bench, Wijk et al., 2024) - 5 tests (2 tests omitted due to security concerns of internet access) **Methodological Details** include (1) Environment and elicitation setup (e.g. METR's modular scaffold with minimal adjustment) (2) Benchmark with human expert and model performance baselines (2) Quantitative evaluation metrics System Card (pp. 17-19); Technical Report (pp. 33-36)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**
Scheming & Misalignment Risks							
Yes **The scope of evaluation** includes alignment faking, undesirable or unexpected goals, hidden goals, deceptive or unfaithful use of reasoning scratchpads, sycophancy toward users, a willingness to sabotage our safeguards, reward seeking, attempts to hide dangerous capabilities, and attempts to manipulate users toward certain views. **Methodology domains** cover the following aspects including: (1) Automated behavioral audits with realism filtering, example seed instructions and evaluation criteria. (2) Third-party replications in collaboration with UK AISI and Apollo Research; (3) Training and pilot-use monitoring that deploys longitudinal checks of whether alignment persists in live internal use and reinforcement-learning logs. (4) Risk-area breakdowns — domain-specific sub-evaluations that map failure modes across self-preservation, sycophancy, reasoning faithfulness etc. (5) White-box interpretability investigations, which is the first public frontier-model effort to examine whether internal representations causally support alignment rather than merely simulating it. The investigations combine unsupervised model diffing and targeted evaluation awareness investigation. System Card (pp. 50-114)	**Yes** Sandbagging is added as part of the Research Categories for the Preparedness Framework. While it does not suffice to be a Tracked Category, it represents a field where "work is required now in order to prepare to effectively address risks of severe harms in the future." **The scope of evaluation** for the broader alignment field includes Sycophancy, Deception (coding deception, browsing broken tools, CharXiv missing image, AbstentionBench), Sandbagging. **Methodology details** include (1) Benchmarks (2) Quantitative evaluation metrics System Card (pp. 7-8, 13-15, 43)	**Yes** **Safety Framework Classification** Deceptive-alignment evaluations are tested Instrumental Reasoning Level 1 and Level 2, both remaining unreached. **Evaluation Scope** covers (1) Stealth evaluations (2) Situational awareness evaluations **Methodological Details** are not included in the model card, but Phuong et al. (2025) publishes the following details: (1) Environment and elicitation setup (e.g .scaffolds that uses chain-of-thought reasoning and reflection to enhance context utilization and better planning) (2) Benchmarks (3) Quantitative evaluation metrics System Card (pp. 19-20)	**Not Mentioned**	**Yes** **The scope of evaluation** includes Manipulation Sycophancy, Deception, and Persuasiveness **Methodology details** include (1) Benchmarks (MASK for deception, OpenAI's MakeMeSay for persuasiveness) (2) Quantitative evaluation metrics System Card (pp. 4, 6)	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**