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
Anthropic has previously run 2 rounds of bug bounty programs in August 2024 and May 2025. Anthropic announced on May 22, 2025, an ongoing bug bounty initiative accepting applications on a rolling basis, as opposed to "invitation-only" in the previous rounds. **Scope:** The program focuses on live deployed systems with ASL-3 protections, and seeks universal and detailed jailbreaks that extract detailed biological-threat information. **Reward:** Up to $35,000 per novel, universal jailbreak identified. (up to $15,000 in August 2024 and up to $25,000 in May 2025) **Timeline:** Issues are resolved usually within ~ 1 days although time to resolution is missing. **Access:** Participants have access to free model aliases that reflect the model and classifiers live on our latest, most advanced model, as opposed to early access to unreleased safety mitigation systems and models in the previous rounds. **Confidentiality:** Formal NDA frameworks	**Scope:** The ongoing bug bounty program covers a wide range of security vulnerabilities across its products and infrastructure, including the OpenAI API, ChatGPT (Plus, plugins, and agent modes), Sora, Atlas. It explicitly excludes model behavior or safety issues (e.g., jailbreaks, hallucinations, prompt content). **Reward** scale by severity: - Critical (P1): up to $100,000 - High (P2): $2,000–$6,500 - Medium (P3): $1,000–$2,000 - Low (P4): $200–$500 **Timeline:** Validation is usually within 6 days. 75% of submissions are accepted or rejected within 6 days in last 3 months. **Access:** Participants test in-scope systems only. API testing, plugin testing (only for self-created plugins), and limited third-party vendor exposure checks are permitted. **Confidentiality:** Partial Safe Harbor for good-faith security research but requires strict confidentiality, prohibiting public disclosure until OpenAI authorizes it (usually within 90 days).	**Scope:** The ongoing AI Vulnerability Reward Program (VRP) covers AI-related security and abuse vulnerabilities in Google/Alphabet AI products, where interaction with an LLM or GenAI system is integral to the bug. Policy or alignment bypasses, jailbreaks, hallucinations, and content violations are explicitly out of scope. Vertex AI and other Google Cloud issues are handled by the separate Cloud VRP. **Reward:** up to US $20,000 for rogue actions detected with flagship products (including Gemini products), adjusting for reporting quality and accounting for novelty bonus (+$1k - +$5k). **Access:** testing limited to researcher's own/test accounts (recommended); no special model access. **Confidentiality:** Participants should follow a designated Code of Conduct, under which they are encouraged to follow coordinated vulnerability disclosure and are expected to have good faith.	**Scope:** The ongoing bug bounty program (started in 2023) is restricted to privacy or security issues, like extracting training data through tactics like model inversion or extraction attacks. (Consistent with the findings of July 2025 AI Safety Index) **Reward:** - The minimum reward for a qualifying submission is US $500. - The maximum reward for a qualifying submission in Meta AI is US $30,000. **Access:** Participants do not have special access to the models but are encouraged to use authorized or test accounts. **Confidentiality:** Meta's Bug Bounty confidentiality and disclosure rules require researchers to avoid privacy violations, use only authorized or test accounts, immediately report and delete any inadvertently accessed data, and give Meta reasonable time to investigate before any public disclosure. Safe-harbor protections apply only if researchers act in good faith and fully comply with these terms.	**Scope:** The program covers xAI, including the Grok API, and targets traditional security vulnerabilities, including authentication, authorization, data-exposure, and infrastructure issues. However, model behaviors and AI safety issues are explicitly out of scope. **Reward:** Bounties are discretionary, determined by a 5×5 internal risk matrix (impact × likelihood) and by a panel of security experts. 90-day averages as of the last update (May 2025): ▪ Low $100 – $500 (19.6 %) ▪ Medium $500 – $2,000 (40 %) ▪ High $2,500 – $7,000 (30 %) ▪ Critical $7,500 – $20,000 (10 %) **Timeline:** Issues are usually triaged within ~1 day and resolved within ~3 weeks. **Access:** No mention of model access or sandbox environment. **Confidentiality:** Participants must abide by HackerOne's disclosure guidelines, including using test accounts, protecting user privacy, and keep all findings confidential until the report is closed.	**Not Mentioned**	**Not Mentioned**	**Not Mentioned**