EU AI Code of Practice Safety and Security	**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
Appendix 3.4-3.5 Signatories must ensure that qualified independent external evaluators assess their models for systemic risk unless the model is already proven comparably safe or evaluators cannot be secured after reasonable efforts. These evaluators must have relevant technical expertise (academic or professional) and follow strict security and confidentiality protocols. Meanwhile, signatories will provide independent external evaluators with (1) adequate access (e.g. access to model activations, gradients, logits, chains-of-thought, model version(s) with the fewest safety mitigations implemented) (2) information (e.g. model specifications (including the system prompt), relevant training data, test sets, and past model evaluation results), (3) time, and (4) other resources (e.g. compute budgets, staffing, engineering budgets and support) Signatories will not undermine the integrity of external model evaluations by storing and/or analyzing inputs and/or outputs from test runs without express permission from the evaluators.	External organizations shared summaries of initial findings for Anthropic to reproduce and compare results with internal investigations for the snapshots and final versions. According to Anthropic's Transparency Hub, "external evaluations use API access with zero-data-retention settings to prevent content storage," consistent with the practices identified in our previous iteration of the AI Safety Index (July 2025). **UK AI Security Institute (UK AISI)** Access: an early snapshot, access released on September 22, 2025) Scope: Misalignment threats (e.g. self-preservation, evaluation awareness etc.) Validation method: Ablations of key environment factors **Apollo Research** Access: pre-deployment snapshot Scope: Misalignment threats (e.g. strategic deceptions, scheming, evaluation awareness etc.) **Independence** (1) Evaluators may publish independently after company review/possible redaction. (2) The company provided its own summary of the evaluator's key findings.	**Scope** SecureBio (Static, agent, and long-form evaluations + manual red teaming for bio risks); Pattern Labs (Evaluates evasion, network attack simulation, and vulnerability discovery and exploitation); METR (AI R&D automation, rogue replication, strategic sabotage); Apollo Research (Covert & deceptive actions); Gray Swan Arena Platform (Prompt-injection and bio-weaponization jailbreaks); FAR.AI (Biological and system-level jailbreak stress tests); U.S. Center on Artificial Intelligence Standards and Innovation (CAISI) (Cyber, biological, and chemical capabilities and safeguards); UK AISI (Cyber and biological / chemical capabilities, plus safeguard penetration testing); Microsoft AI Red Team (Frontier Harms, Content Safety, and Psychosocial Harms). **Access** (1) The longest period of time that an external evaluator was given continuous access for pre-deployment testing is >2 weeks (<=3 weeks). (2) The highest level of technical access granted to any of the listed external evaluators is Standard inference API with normal user-facing filters in place, Inference API with safety filters disabled (no inference-time mitigations), and "Helpful-only" or base model API (no harmlessness fine-tuning and no filters). (3) Third party assessors were provided OpenAI GPT-5 Thinking early checkpoints, as well as the final launch candidate models. **Security** Zero Data Retention available upon request, if technically feasible during pre-deployment periods **Independence** (1) Evaluators may publish independently without prior company approval after the model is released, provided that evaluations are run independently on the deployed model. (2) Evaluators may publish independently after company review/possible redaction. Since pre-deployment evaluation period are under NDA, publications require prior approval to protect confidential information. METR has published the full report. (3) The company provided its own summary of the evaluator's key findings, which they share with the evaluator prior to release to confirm factual accuracy. (4) OpenAI publishes excerpts from the report mutually agreed upon or written, with the company having the final say for what content goes in System Cards. **Timeline** External safety tests were completed after broad internal deployment.	External organizations are chosen based upon their domain expertise, and include civil society and commercial organizations. However, they are not named individually. **Scope:** Autonomous systems, cybersecurity, CBRN, and societal risk **Access:** (1) The highest level of technical access granted to any of the external evaluators is the Black-box access to Gemini 2.5 Pro (Preview 05-06) via the inference API, with safety filters disabled (no inference-time mitigations). (2) The longest period of time that an eternal evaluator was given continuous access for pre-deployment is >3 weeks (<=5 weeks). (3) For pre-deployment testing, evaluators had higher quotas for query rates than the public/enterprise tier but were still subject to explicit caps (e.g. requests-per-minute or daily token limits). The quota is bespoke depending on the testing partner's specific needs and evaluation type. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks. **Independence:** These organizations are independent in choosing methodologies, ranging from qualitative red-teaming to quantitative automated testing, at varying time commitments. After receiving all analyses, raw data, and evaluation materials, internal experts reviewed model outputs and applied harm-severity ratings under established safety frameworks and Critical Capability Levels, and writing reports internally. External evaluators are financially compensated by Google DeepMind for their time. Technical Report (pp. 36-38), Company Survey	**Not Mentioned**	xAI has responded that external testing was commissioned in the survey response without naming the evaluators. The external safety tests were completed before broad internal deployment. They released the same model version that the final round of safety evaluations were conducted on. **Access:** The highest level of technical access it has shared externally is Helpful-only' or base model API (no harmlessness fine-tuning and no filters), with the longest duration of more than 5 weeks. Evaluators will have higher quotas than the public or enterprise tiers for query rates but are still subject to explicit caps (e.g. requests-per-minute or daily token limits. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. **Independence:** Evaluators may publish independently after company review or possible redaction. **Timeline:** All external safety tests were completed before broad internal deployment. Source: Company Survey	**Not Mentioned**	**Scope:** Z.ai has collaborated with China Academy of Information and Communications Technology (CAICT), which is a subordinate to the powerful Ministry of Industry and Information Technology (MIIT), for evaluations of "general safety issues," as according to the survey response. **Access:** The highest level of technical access it has shared externally is 'Helpful-only' or base model API (no harmlessness fine-tuning and no filters). There are no limits for query-rate or volume restrictions to external evaluators. **Security:** Inputs and outputs are neither logged nor retained, protecting evaluator IP. **Independence:** Evaluators may publish independently without prior company approval after the model is released. **Timeline:** All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment. Source: Company Survey	**Not Mentioned**