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 3.2 Signatories are required to conduct model evaluations using at least state-of-the-art elicitation methods that minimize under-elicitation and model deception during model evaluation, and that match both the capabilities of potential misuse actors and the model's expected use context. Examples of the measures include adapting test-time compute, rate limits, scaffolding, tools, fine-tuning, and prompt engineering	**Adapting test-time compute** is reported in cyber evaluations (e.g. flexible token constraints) and CBRN evaluations (e.g. pass@5 results reported for longform virology, extended thinking) and alignment evaluations (extended thinking) **Scaffolding** is reported in cyber evaluations (e.g. specific resets and auto-summorization in CyberGym) **Iterative Prompting** is reported in CBRN evaluations (e.g. prompt engineering based on analyzing failure cases) **Tool use** is reported in CBRN evaluations (e.g. tools and agentic harnesses) and cyber evaluations (e.g. code editor and a terminal tool) **Helpful-only variants** are reported in CBRN evaluations.	**Adapting test-time compute** is reported in cyber evaluations (e.g. pass@12 for CTF challenges and cyber range evaluations) and and AI self-improvement evaluations (e.g. SWE-bench and MLE-Bench multi-rollout trials). **Tool use** is reported in cyber evaluations. **Custom post-training** (e.g. helpful-only variants), **scaffolding** and **prompting** are applied where relevant, though the System Card does not specify which evaluations each technique was used in.	**Scaffolding and Agent Harness** is reported in cybersecurity, machine-learning R&D, and deceptive-alignment tests, which includes chain-of-thought and reflection loops. **Tool use** is reported in cybersecurity evaluations. **Parallel attempt setups** is reported by cybersecurity evaluations (10-50 attempts) and deceptive-alignment tests (50 retries) and meanwhile time and run budgets (43 × 45-minute vs 16 × 2-hour runs) are mentioned for ML R&D benchmarks. **Prompt engineering** is reported in CBRN and cybersecurity (e.g. open-ended, multi-turn).	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned	Not Mentioned