

xAI publishes biosecurity results for Grok 4.6 from an outside evaluator

Source copy kept by wheresthe.ai for the story "xAI publishes biosecurity results for Grok 4.6 from an outside evaluator" (<https://wheresthe.ai/d/xai-biosecurity-at-the-frontier-2026-09/>).

Source: <https://x.ai/news/biosafety-at-the-frontier>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 91e3e4dcdf8a864c4a2760a361f325d5ebdf25b100e43813551c62efe7e0bdd2

Documents the source offers: <https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf>

Biosecurity at the frontier | SpaceXSI Back to news Sep 1, 2026 Biosecurity at the frontier

Today, LatchBio published an independent analysis of Grok 4.6 on their biological capability and biological red-teaming benchmark suites. These evaluations measure a wide range of characteristics and capabilities of tested models, including:

Whether a model can distinguish between legitimate biological research and requests that conceal a biosecurity hazard

Whether a model is useful in a biosecurity and biomonitoring role, by being able to detect and characterize emerging pathogens and patterns of transmission

On LatchBio's BioSecBench-Refusal suite, Grok 4.6 was the strongest model tested at refusing disguised and hazardous tasks while still completing routine biological work. It was the only system to score above 50% on both measures. Methodology and context Biology and biological research present domains where the opportunities and risks of highly capable AI systems are well-known and challenging to disentangle. In designing our safeguards, we aim to maximize the utility of our agents in assisting and accelerating legitimate scientific research, while minimizing the risks they present through incorrect information and adversarial use. LatchBio's two benchmarks most relevant to the above capabilities measure whether an agent can distinguish concealed hazards from ordinary science, and whether it can carry out the pathogen surveillance workflows public-health work depends on. They are:

BioSecBench-Refusal , which pairs routine biological tasks, adapted from published literature, with 46 red-team tasks that look like ordinary research and hide the hazard in attached scientific data, mislabeled files, or other intentionally obfuscated tasks. This evaluation is calibrated around understanding the intent of a task: an agent that only reacts to words like "pathogen" or "toxin" will block the routine set and miss the red-team set, while one that inspects the data can tell the two apart.

BioSecBench-Surveillance , which tests whether an agent can carry out pathogen genomic surveillance workflows of the kind used in public-health monitoring. Those tasks require chaining file inspection, tool use, and scientific judgment on messy sequencing data.

Results and key findings In testing on BioSecBench-Refusal, LatchBio found that Grok 4.6 detects and refuses red-team and otherwise dangerous queries more reliably than any other frontier system tested, while not diminishing performance on routine biological work. Grok 4.6 also performs comparably with other frontier models on biosurveillance work. LatchBio notes that evaluations were performed on a variety of agent harnesses to remove confounds, and that unless otherwise noted, agents were tested at their highest-offered effort levels. The score presented by LatchBio is a trial-weighted harmonic mean of red-team refusal and routine compliance. We analyze this metric and standalone refusal rate independently. Across different harnesses, Grok 4.6 holds the top three spots, averaging 62.1%. Considering refusals and task compliance independently, Grok 4.6 refused 59.2% of red-team tasks and completed 64.8% of routine ones. It is the only model tested that scored above 50% on both measures. On BioSecBench-Surveillance, Grok 4.6 averages a success rate of 53.5%, sitting behind Opus 5 and ahead of GPT-5.6 Sol on biosecurity and monitoring work. The results on both benchmarks indicate an agent and underlying model well-calibrated for routine and helpful biological work, as well as one highly capable in biosecurity work and monitoring. On evaluations of general routine biological capability performed by LatchBio (such as SpatialBench or TxBench-PP), Grok is found to match or exceed other frontier models in a wide range of agentic biological work. These are presented in greater detail at benchmarks.bio . Analyzing refusal behavior In

evaluation traces, Grok 4.6 is observed reasoning over the contents of a task and testing environment to assess intent before proceeding or refusing. Frequently, Grok will find discrepancies between the stated intent in the prompt and environment, or will assemble intent from high-risk content disguised by filenames and encryption, and will subsequently refuse. On obviously benign and low-risk tasks, Grok exhibits the same environment-reasoning behavior, but is able to assess tasks as safe. How we safeguard Grok Before we release a model, we test biological capability in addition to other domains of risk, and validate that the capabilities of our models are adequately safeguarded against misuse. The work of third-party evaluators such as LatchBio complements the evaluations we perform internally, both pre- and post-deployment, for the models we serve. Grok's safeguards are built in layers to establish defense-in-depth. Refusal training is performed to teach the model how and when to refuse, to correctly infer the intent and risk profile of tasks, and to refuse correctly in highly adversarial scenarios. We train and deploy inference-time safeguards to reject harmful requests before they ever reach the model, and we implement behavioral controls to further safeguard the model when deployed. Post-deployment monitoring is performed to detect and stop patterns of adversarial use at the session and user level, and provides a source of feedback for continuous calibration of our model deployments. We observe a material improvement in Grok 4.6's capabilities in biological work over historically tested models, with substantial gains in refusal and biosecurity performance over Grok 4.5 and Grok 4.3. We intend our ongoing and novel safeguards work to allow us to continue serving frontier-scale intelligence safely across future model releases. The future of biosecurity and scientific discovery The line between actively adversarial tasks and helpful use grows thinner as models rapidly become more capable and autonomous, and safeguards will need to evolve accordingly. To control for the greater risk implied by improved model capability and agency, we will perform wider, more ambitious testing and calibration of our models and agentic systems: broader pre-deployment suites, more third-party evaluations, improved post-deployment monitoring, and deployments of our models with companies and institutions at the frontier of biology. Similarly, there is a risk inherent in overrefusals and miscalibrated safeguards. When a model refuses routine and helpful biological work, the ability of healthcare professionals, researchers, and monitoring programs to detect outbreaks early and perform other critical work in the field is degraded. We gauge this risk as equally serious as the risk of aiding malicious use. Grok is already used in scientific work, including general biological research and biosecurity monitoring, and we are optimistic about the ability of highly capable models to rapidly accelerate scientific discovery and shorten the path from the lab to practical use. We are committed to continuing to serve frontier intelligence securely and safely to the engineers, researchers, enterprises, and institutions advancing the life sciences. LatchBio's full methods and per-model scores are presented at [benchmarks.bio](#) . LatchBio's complementary blog, the Grok 4.6 model card, and SpaceXSI's Frontier Artificial Intelligence Framework are linked below.

[LatchBio blog](#)

[Grok 4.6 model card](#)

[Frontier Artificial Intelligence Framework](#)