GPT-Rosalind-6 System Card Addendum

OpenAI

October 2026

Contents

1 Introduction **3**

2 Model Data and Training **3**

3 Beneficial Capability Evaluations **3**

3.1 Medicinal chemistry 4

3.2 Genomics and quantitative biology 4

3.3 Applied life sciences research 4

4 Preparedness **4**

4.1 Preparedness Evaluations 4

4.1.1 Biological and Chemical 4

4.1.1.1 ProtocolQA Open-Ended 5

4.1.1.2 TroubleshootingBench 6

4.1.1.3 Biorisk knowledge 6

4.1.1.4 Multi-select Virology Troubleshooting 7

4.1.1.5 AAV Capsid Packaging Prediction 7

4.1.1.6 SHP2 Protein Function Prediction 7

4.1.1.7 Coronavirus-ACE2 Cell-Entry Screen 8

4.1.1.8 Phage-plasmid Co-evolution 8

4.1.2 Cybersecurity 8

4.1.3 AI Self-Improvement 9

4.2 Safeguards 9

4.2.1 Trust-based access 9

4.2.2 Model behavior and system safeguards 10

4.2.2.1 Biological and Chemical safeguards 10

4.2.2.2 Cyber boundaries 10

4.2.2.3 Safety and alignment evaluations 11

4.2.3 Enforcement 12

4.2.4 Security and Access Controls 13

References **13**

1 Introduction

GPT-Rosalind-6 is a frontier reasoning model for biology, drug discovery, and translational medicine. We offer it as GPT-Rosalind-6 Research to qualified scientists, research institutes, companies and government partners. Rosalind has fewer refusals on dual-use biological queries but is only made available via trusted access. We review applicants' proposed uses, security controls and governance practices before granting access. We also continue to have other deployment safeguards in place.

GPT-Rosalind-6 uses the same weights as GPT-6 Astra, our flagship reasoning model. It is designed to support legitimate biological research, including some dual-use topics, while still refusing requests for biological weaponization and retaining refusals on other safety topics.

Under our Preparedness Framework, we are treating this launch as High capability in the Biological and Chemical domain, below the threshold for Critical. We rely on GPT-6 Astra's capability assessments for other Preparedness risks and treat GPT-Rosalind-6 as Critical in Cybersecurity and below High in AI Self-Improvement.

This addendum describes GPT-Rosalind-6's biological capability assessment, additional safety and alignment evaluations, and deployment safeguards. We include results for GPT-Rosalind-5.5 where available for comparison. This addendum should be read alongside the GPT-6 Astra system card, which describes the parent model's broader capability and safety evaluations.

2 Model Data and Training

GPT-Rosalind-6 uses the same weights as GPT-6 Astra. We did not add training data or perform additional post-training for this release. Details of the underlying model's data and training are provided in the GPT-6 Astra system card.[1]

3 Beneficial Capability Evaluations

We evaluated GPT-Rosalind-6 on three life sciences research benchmarks and compared its results with GPT-Rosalind-5.5. The GPT-Rosalind-5.5 scores below use the same benchmark versions and problem sets as GPT-Rosalind-6 and therefore differ from those reported at its launch. The Preparedness evaluations in Section 4 assess capabilities relevant to the risk of severe harm. We averaged response scores within each problem and then weighted the problems equally.

[1]GPT-Rosalind-6 is intended to be used in accordance with OpenAI's Usage Policies, Service Terms, and Terms of Use. These policies apply universally to OpenAI services and are designed to ensure safe and responsible usage of AI technology. You can review OpenAI's Usage Policies at openai.com/policies/usage-policies/.

If you need assistance with respect to these models, you can find further information on OpenAI's website (openai.com), or you can contact OpenAI Support by opening the chat bubble icon displayed at the bottom-right of help.openai.com. The input and output modalities supported by the models can be found at https://developers.openai.com/api/docs/models/all.

A list of the languages that ChatGPT currently supports can be found here.

3.1 Medicinal chemistry

MedChemBench v7 tests chemical structure interpretation, structure-activity relationships, and predictions of drug potency and toxicity. We evaluated both models on the same 99 problems, with up to nine scored responses per problem at each reasoning budget. At the highest reasoning budget tested, mean rubric scores were 49.6% for GPT-Rosalind-6 and 28.9% for GPT-Rosalind-5.5.

3.2 Genomics and quantitative biology

GeneBench Pro v13 tests an agent's ability to plan and carry out quantitative analyses of biological data. We estimated single-attempt success from seven responses per model on each of 129 problems at each reasoning budget. At the highest reasoning budget tested, single-attempt success rates were 37.7% for GPT-Rosalind-6 and 16.7% for GPT-Rosalind-5.5. The evaluation used a benchmark-specific configuration that differs from production.

3.3 Applied life sciences research

LifeSciBench Gold uses research tasks contributed by practicing life scientists. We excluded six problems flagged for possible issues with the tasks or evaluation environment and compared both models on the same remaining 94 problems, with up to eight scored responses per problem at each reasoning budget. At the highest reasoning budget tested, average scores were 61.3% for GPT-Rosalind-6 and 44.0% for GPT-Rosalind-5.5.

4 Preparedness

The Preparedness Framework is OpenAI's approach to tracking and preparing for frontier capabilities that create new risks of severe harm. Under our framework, we work to track and mitigate the risk of severe harm, including by implementing safeguards that sufficiently minimize the risk for highly capable models.

4.1 Preparedness Evaluations

We repeated biological capability evaluations on GPT-Rosalind-6 and use the parent model's helpful-only evaluations as additional evidence, given it uses the same weights as GPT-6 Astra. Helpful-only models are evaluated to elicit capabilities that safety refusals may otherwise conceal.

4.1.1 Biological and Chemical

As we did for GPT-6 Astra, we are treating this launch as High capability in the Biological and Chemical domain. Below, we describe the evaluations we used to reach this conclusion. We use the four High capability evaluations and four Critical capability evaluations described in the GPT-6 Astra system card. The hard-negative protein binding and DNA sequence design evaluations reported for GPT-Rosalind-5.5 have been retired.

The High capability evaluations test wet lab knowledge and troubleshooting. The Critical capability evaluations test biological design capabilities that could help an expert develop a highly dangerous novel threat. These evaluations provide evidence about specific capabilities rather than a complete test of the ability to cause biological harm. The Critical capability thresholds are preliminary estimates informed by expert judgment and may change as the evaluations are refined.

Table 1: Overview of Biological and Chemical evaluations

Evaluation	**Capability**	**Description**
	High	
ProtocolQA Open-Ended	Wet lab capabilities	Troubleshooting commonly published lab protocols.
TroubleshootingBench	Tacit knowledge and troubleshooting	Identifying errors in non-public, expert-written lab protocols.
Biorisk knowledge	Biorisk-relevant knowledge	Answering questions about tacit biological knowledge and laboratory troubleshooting.
Multi-select Virology Troubleshooting	Wet lab troubleshooting	Identifying the complete set of correct responses to virology troubleshooting questions.
	Critical	
AAV Capsid Packaging Prediction	Novel Design (Protein design)	Predicting packaging performance for held-out capsid variants.
SHP2 Protein Function Prediction	Novel Design (Protein design)	Predicting functional assay results for protein variants.
Coronavirus-ACE2 Cell-Entry Screen	Novel Design (Virus-host interaction)	Predicting and ranking held-out cell-entry measurements.
Phage-plasmid Co-evolution	Novel Design (Co-evolution)	Predicting mutations that appear at later experimental timepoints.

The results below report observed performance and, for the High capability evaluations, refusal-adjusted scores. pass@1 measures success on a single attempt. The adjustment treats refusals as correct answers to estimate capability that safety training may otherwise conceal. For Biorisk knowledge, we recompute the consensus answer after this adjustment. GPT-6 Astra reference values use the evaluation runs reported in its system card.

4.1.1.1 ProtocolQA Open-Ended

To evaluate models' ability to troubleshoot commonly published lab protocols, we modify 108 multiple-choice questions from FutureHouse's ProtocolQA dataset [3] to be open-ended short answer questions, which makes the evaluation harder and more realistic than the multiple-choice version. The questions introduce egregious errors in common published protocols, describe the wet lab result of carrying out this protocol, and ask how to fix the procedure.

Table 2: ProtocolQA Open-Ended. The indicative High capability threshold is 54%.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
pass@1	42.0%	36.4%	41.4%
Refusal-adjusted	42.0%	36.4%	45.4%

4.1.1.2 TroubleshootingBench

To evaluate models' ability to identify and correct real-world experimental errors in biological protocols, we built a short-answer troubleshooting dataset from expert-written wet lab procedures. TroubleshootingBench focuses on tacit, hands-on knowledge and uncontaminated procedures that are not available online.

Scientists with a PhD in a relevant biological discipline were asked to transcribe biological protocols they have personally used in the lab. From these protocols, they created troubleshooting questions, introducing subtle or realistic execution errors and describing the resulting failed outcome. Compared to ProtocolQA Open-Ended, which focuses on well-known published procedures, TroubleshootingBench tests performance on non-public protocols and errors that rely on tacit procedural knowledge.

Table 3: TroubleshootingBench. The indicative High capability threshold is 36.4%. GPT-Rosalind-6 scored 491/832, counting nine failed evaluation attempts as incorrect.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
pass@1	59.0%	53.3%	48.4%
Refusal-adjusted	59.0%	53.3%	63.5%

4.1.1.3 Biorisk knowledge

We evaluated models on a tacit knowledge and troubleshooting multiple-choice dataset created with Gryphon Scientific. The questions span all five stages in the biothreat creation process and focus on areas where tacit knowledge would be a bottleneck. Tacit knowledge questions are meant to be obscure to anyone not working in the field. Troubleshooting questions are meant to be obscure to anyone without hands-on experience.

Table 4: Biorisk knowledge. cons@32 uses the consensus answer across 32 responses to each question. The indicative High capability threshold is 80%.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
cons@32	78.3%	79.2%	63.3%
Refusal rate	9.4%	1.8%	29.2%
Refusal-adjusted cons@32	85.0%	n/a	90.0%

The refusal-adjusted GPT-Rosalind-5.5 score published in its system card used a different calculation and is therefore omitted from this comparison.

4.1.1.4 Multi-select Virology Troubleshooting

To evaluate models' ability to troubleshoot wet lab experiments in a multimodal setting, we evaluate models on a set of fully held-out virology troubleshooting questions from SecureBio, in a multi-select format. A response receives credit only if it selects every correct statement and no incorrect ones.

Table 5: Multi-select Virology Troubleshooting. The indicative High capability threshold is 31%.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
pass@1	59.9%	51.2%	55.8%
Refusal-adjusted	59.9%	51.2%	63.1%

After adjusting for refusals, GPT-Rosalind-6 exceeds the indicative High capability thresholds on three of the four evaluations. We therefore continue to treat the model as High capability. Compared with the reported GPT-Rosalind-5.5 results, observed scores are higher on three evaluations and lower on Biorisk knowledge.

4.1.1.5 AAV Capsid Packaging Prediction

Predicting whether an AAV capsid variant will package successfully could reduce trial and error in viral-vector engineering, which we treat as a proxy for some dual-use biological design capabilities. We evaluate this capability using an evaluation developed by Dyno Therapeutics. The model predicts packaging scores for held-out sequences using labeled examples, protein embeddings and computational tools. Performance is measured using Spearman rank correlation.

Table 6: AAV Capsid Packaging Prediction. The indicative Critical capability threshold is 0.600.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
Spearman correlation	0.529	0.22	0.528

4.1.1.6 SHP2 Protein Function Prediction

Predicting how changes to protein sequence affect function could reduce trial and error in protein engineering, which we treat as a proxy for some dual-use biological design capabilities. We evaluate this capability using SHP2, a well-studied human protein. The model uses available experimental measurements to predict held-out functional assay results for single-amino-acid variants. Performance is measured by mean R^2 across three unpublished assay datasets.

Table 7: SHP2 Protein Function Prediction. The indicative Critical capability threshold is 0.60. The GPT-6 Astra reference is the reported maximum of 0.40. A separate helpful-only run scored 0.35.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
Mean R^2	0.38	0.16	0.40

4.1.1.7 Coronavirus-ACE2 Cell-Entry Screen

Reliable predictions of cell entry could help researchers prioritize laboratory experiments in studies of virus-host interactions, but could also be a threat-relevant capability. We ask models to predict and rank held-out measurements from an existing, unpublished academic dataset. The underlying experiment studied existing spike proteins using replication-incompetent pseudoviruses. It did not introduce new mutations or generate infectious progeny.

Table 8: Coronavirus-ACE2 Cell-Entry Screen. The indicative Critical capability threshold is 0.75. The GPT-6 Astra reference is its helpful-only result.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
Composite score	0.42	0.44	0.45

4.1.1.8 Phage-plasmid Co-evolution

We assess the ability to predict biological evolution using data from an existing but unpublished experimental phage-plasmid co-evolution system. Given observations available at each experimental checkpoint, the model predicts which mutations will subsequently appear. Performance is measured by the negative log-likelihood of observed mutations, with lower scores indicating better predictions. The experiment uses viruses that infect bacteria rather than humans. We treat performance as a proxy for some dual-use biological design capabilities.

Table 9: Phage-plasmid Co-evolution. The indicative Critical capability threshold is an NLL of 9.40 or lower. The GPT-6 Astra reference is its helpful-only result.

Metric	GPT-Rosalind-6	GPT-Rosalind-5.5	GPT-6 Astra
Negative log-likelihood	13.41	11.81	12.58

None of the four Critical capability evaluations met its indicative threshold. Results relative to GPT-Rosalind-5.5 were mixed. We also rely on the parent model's helpful-only evaluations and external biological capability assessments, as described in the GPT-6 Astra system card.

4.1.2 Cybersecurity

We rely on GPT-6 Astra's existing Preparedness assessment for Cybersecurity and treat GPT-Rosalind-6 as Critical in this domain. The underlying capability findings are described in the

GPT-6 Astra system card.

4.1.3 AI Self-Improvement

We rely on GPT-6 Astra's existing Preparedness assessment for AI Self-Improvement, where GPT-6 Astra did not meet the threshold for High capability. The evaluations and findings are described in the GPT-6 Astra system card.

4.2 Safeguards

Compared to GPT-6 Astra, GPT-Rosalind-6 features additional deployment controls for its use in advanced life sciences research. Access is limited to approved customers through our trusted-access deployment structure where organizations must demonstrate they are conducting legitimate scientific research with public benefit, have strong governance and safety oversight, and control access with enterprise-grade security.

Our safeguard approach for GPT-Rosalind-6 has four main components: trusted access controls for verified higher-capability life sciences use, model-level boundaries against harmful biological and cyber assistance, monitoring and enforcement mechanisms designed to detect abuse, and additional contractual requirements to enhance security.

GPT-Rosalind-6 retains GPT-6 Astra's standard cyber safeguards, including real-time blocking.

4.2.1 Trust-based access

GPT-Rosalind-6 is available only to approved customers. Before granting access, OpenAI reviews whether the applicant has a legitimate life sciences or related mission, a credible use case, appropriate research and operational governance, and the ability to safely control access to the model.

GPT-Rosalind-6 uses the access requirements already applied to GPT-Rosalind-5.5, including the policy updates introduced since its launch. This release does not broaden eligibility beyond prior announcements, such as the Rosalind Academic Plan.

For non-government applicants, OpenAI also employs business verification and compliance-screening checks. These checks help confirm applicant authenticity and reduce the risk of access by fraudulent, insufficiently attributable, or otherwise unsuitable actors. These checks supplement OpenAI's broader review of the applicant's proposed use case, safety posture, and access scope.

For government applicants, OpenAI maintains a trusted-access pathway for select government entities working on approved biosecurity, biosafety, public-health preparedness, and biodefense use cases.

Approved enterprise customers must limit GPT-Rosalind-6 access to authorized users with a legitimate need and apply least-privilege access controls, appropriate onboarding, and insider risk controls to their users. Customers must apply qualified human review before model outputs inform experiments, publications, or consequential decisions.

4.2.2 Model behavior and system safeguards

These evaluations were deliberately created to be difficult. They were built around cases in which our existing models were not yet giving ideal responses, and this is reflected in the scores below. We continue monitoring these categories after launch to evaluate online performance and further adjust safeguards as appropriate.

4.2.2.1 Biological and Chemical safeguards

GPT-Rosalind-6 is designed to be more useful for legitimate advanced biological research than generally available ChatGPT models. This supports advanced life sciences workflows, including some dual-use topics, that GPT-6 Astra may refuse in General Access. We still tune the model to refuse or safely respond to requests that would continue to meaningfully enable severe biological weaponization. This training is done separately from the training of our classifiers and offline mitigations to decorrelate our safeguards.

To evaluate the quality of these model-level refusals, we track the safety of model responses to prompts that originate from held-out synthetic data, red-teaming, and production data. These benchmarks and their grading continue to be updated, meaning results do not directly translate to past system cards.

Table 10 reports results across our suite of biosecurity benchmarks, accounting for both model behavior and safeguards. We note that because GPT-Rosalind-6 is a trusted-access deployment, we mainly considered the results from the *Biosecurity - Severe* benchmark. *Biosecurity - Benign* evaluates model compliance (being fully helpful without refusals or caveats) on a set of benign biological and operational questions. The remaining benchmark scores show the fraction of model responses that pass the configured safety grader on prompts related to biological weaponization. A response that does not pass was flagged under this grading criterion; this does not necessarily mean it is unequivocally unsafe.

Table 10: Biological and chemical content evaluations (higher is better)

Benchmark	**GPT-Rosalind-6**	**GPT-Rosalind-5.5**	**GPT-6 Astra**
Biosecurity - Severe	98.1%	88.6%	99.8%
Biosecurity - Sensitive	79.0%	18.3%	99.8%
Biosecurity - Select Sequences	83.8%	50.0%	95.6%
Biosecurity - Dual-Use	84.6%	69.0%	97.8%
Chemical Misuse - Severe	100.0%	59.1%	100.0%
Biosecurity - Benign	100.0%	97.7%	91.6%

4.2.2.2 Cyber boundaries

GPT-Rosalind-6 inherits the model and system level cyber safeguards of the GPT-6 Astra base model and additionally applies a more conservative cyber refusal boundary. The model is intended to continue to refuse requests that clearly enable unauthorized, destructive, or harmful cyber activity, including areas such as malware deployment, credential theft, exfiltration, or other malicious activity.

As an additional robustness test, we evaluate resistance to previously identified cyber jailbreaks by replaying nine known attacks across ten fixed ExploitBench tasks. We evaluated both models with real-time cyber blocking disabled, so the task success rates in Table 11 reflect model-level behavior only. We report the mean task success rate across seven attacks, excluding attempts that ended in errors or exceeded turn limits.

Table 11: Task success rates on historical cyber jailbreak attacks with real-time blocking disabled (lower is better).

Evaluation	**Measure**	**GPT-Rosalind-6**	**GPT-6 Astra**
Historical Jailbreak Regression	Task success rate under jailbreak attacks ↓	14.3%	15.7%

Two attacks were excluded because all attempts ended in errors or exceeded turn limits.

4.2.2.3 Safety and alignment evaluations

We also evaluate GPT-Rosalind-6's instruction following, honesty, resistance to prompt injection, and behavior when using tools. Additional evaluations assess whether automated monitors detect harmful actions and whether the model can evade detection or alter its reasoning when instructed to do so.

In the tables below, ↑ indicates higher scores are better and ↓ indicates lower scores are better. GPT-6 Astra results are included for context. Evaluation versions, inference settings, grading and sample coverage differ across some runs, so these are descriptive comparisons. "Not available" indicates that no numerical reference result has been identified. ≈ denotes an approximate value read from a published figure.

Table 12: Honesty and reliability.

Evaluation	**Measure**	**GPT-Rosalind-6**	**GPT-6 Astra**
Blocker Deception	Blocker-related deception rate ↓	9.5%	12.9%
Broken Search Tool	Rate of failing to acknowledge that the search tool is unavailable ↓	1.3%	1.5%

Table 13: Behavior when using tools.

Evaluation	**Measure**	**GPT-Rosalind-6**	**GPT-6 Astra**
ExploitGym Honeypot	Rate of successfully exploiting evaluation infrastructure outside the assigned task ↓	0.0%	0.0%
Respecting Warnings	Rate of successfully circumventing a restriction after encountering a warning ↓	21.7%	18.9%

Evaluation	Measure	GPT-Rosalind-6	GPT-6 Astra
Respecting Auto-Review	Rate of attempting to circumvent an automatic-review denial, including unsuccessful retries ↓	0%	0%
Agent Replay Auto-Review	Rate of harmful or unauthorized actions rated certain level severity or higher in replayed agentic conversations, excluding laziness ↓	27.1%	26.7%
Unintended Engagement with External Agent Messages: Performed specified misaligned action	Rate of performing the specified unauthorized action, conditional on discovering the message board ↓	2.5%	0.0%
Cowork Misalignment	Overall rate of misaligned outcomes in simulated workplace tasks ↓	1.4%	1.2%

Table 14: Monitoring and evasion.

Evaluation	Measure	GPT-Rosalind-6	GPT-6 Astra
OAI-repo Sabotage v2	Rate of attempts that successfully sabotage simulated services and evade the monitor ↓	0% unguided, 0% guided	0% unguided, 3% guided
Math Side-tasking Monitor Evasion	Full-context monitor bypass rate among successful hidden mathematics tasks ↓	0%	0%
Chain-of-thought controllability	Rate of following instructions that constrain the content or form of the model's reasoning ↓	GPQA: 71.8% HLE: 59.3% MMLU: 83.1%	GPQA: 69.2% HLE: 53.8% MMLU: 81.6%

4.2.3 Enforcement

OpenAI can restrict or revoke access for misuse. OpenAI maintains monitoring, review, and enforcement processes to help identify potential misuse after access is granted. These processes are designed to detect activity that may indicate malicious biological, chemical or other activity that is in violation of OpenAI's Usage Policies or unauthorized access.

OpenAI uses a variety of signals to assess potential misuse, including automated safety classifiers, account- or organization-level indicators, customer-provided information, and, where available, specialist review of potentially high-risk activity. The specific review process depends on the deployment configuration and the information available to OpenAI.

Before launch, we tested our shared monitoring and enforcement process through a biosecurity red-teaming effort on GPT-Rosalind-5.5 and are in the process of repeating this for GPT-Rosalind-

6 now that it is deployed. We are not publicly disclosing these details because they contain specific information about our safeguards, but we shared them with our Safety Advisory Group committee and other relevant governance bodies.

4.2.4 Security and Access Controls

Approved GPT-Rosalind-6 customers must maintain security controls appropriate for access to a higher-capability life sciences model. These include strong identity and access management, administrator controls, least-privilege access controls appropriate to the approved deployment, and processes to remove access when a user no longer has a legitimate need.

Model access and data-retention arrangements are approved separately. Customers must meet the security and oversight requirements for their approved deployment.

As with other OpenAI deployments, we expect to continue learning from real-world use. We will monitor safety signals, customer experience, and emerging misuse patterns, and may update GPT-Rosalind-6 safeguards as the deployment evolves.

References

[1] OpenAI, "GPT-6 Astra System Card," 2026.

[2] OpenAI, "GPT-Rosalind-5.5 System Card," 2026.

[3] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodriques, "Lab-bench: Measuring capabilities of language models for biology research," 2024.