GPT-Rosalind-5.5 System Card

OpenAI

June 3, 2026

Contents

1 Introduction

GPT-Rosalind-5.5 is a new model in the GPT-Rosalind series, our frontier reasoning model built to support research across biology, drug discovery, and translational medicine. We are deploying it in research preview to trusted organizations, providing access limited to qualified scientists, research institutes and government partners who are working on beneficial uses and who have a strong security and governance posture.

GPT-Rosalind-5.5 is incrementally trained from GPT-5.5, our flagship reasoning model. Outside the biological domain, its capabilities are comparable to GPT-5.5 and it receives the same model-level safety training. We therefore have not provided evaluation results from those capabilities in this system card.

We worked with industry experts to design and evaluate GPT-Rosalind-5.5 on a series of benchmarks across medicinal chemistry, quantitative biology, wet lab protocol assistance, and applied life sciences research. Beneficial biological capabilities, as expected, exceed those of the regular GPT-5.5 model, often while expending fewer tokens.

Under our Preparedness Framework, the Preparedness evaluations in the Biological and Chemical domain met our threshold for High capability while falling below the threshold for Critical. Cybersecurity capabilities were found to be at or below the level of GPT-5.5. Because GPT-Rosalind-5.5 was trained on additional biology data unrelated to AI self-improvement, we rely on GPT-5.5's existing Preparedness assessment for AI Self-Improvement, where GPT-5.5 did not meet the threshold for High capability. Our Safety Advisory Group approved the safeguards plan for this release, finding that it sufficiently mitigates the associated risks of severe harm.

Below we describe the incremental biological capability findings for GPT-Rosalind-5.5 compared with GPT-5.5, and describe the differences in safeguard approach for this deployment compared with the primary GPT-5.5 model.

2 Model Data and Training

GPT-Rosalind-5.5 is incrementally trained from GPT-5.5 with training data related to beneficial life sciences capability. Unlike GPT-5.5, it is trained not to refuse sophisticated biology queries, and leverages a trusted access and responsible deployment structure as the primary safeguard.

3 Beneficial Capability Evaluations

The Beneficial Capability Evaluations below measure how well GPT-Rosalind-5.5 performs on the intended research-support tasks it was designed to improve. Preparedness evaluations, in the following section, measure related but distinct risk-relevant capabilities under OpenAI's Preparedness Framework, where results are interpreted against severe-harm thresholds to inform safeguards and deployment decisions, rather than as the primary measure of product benefit.

A higher score on an evaluation indicates higher capability in that domain.

3.1 Medicinal chemistry

To evaluate GPT-Rosalind-5.5 on medicinal chemistry, a field focused on turning molecules into useful drugs, we designed a multimodal evaluation designed to measure realistic medicinal chemistry workflows, evaluating multimodal chemical structure understanding; structure-activity relationship (SAR); prediction of drug potency, toxicity, and absorption, distribution, metabolism, excretion (ADME); multiparameter lead-optimization decision-making, and retrosynthesis.

Table 1

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**
pass@1	27.5%	25.1%
Average output tokens	27133.5	29231.8

3.2 Genomics and quantitative biology

To evaluate GPT-Rosalind-5.5 on genomics understanding and quantitative biology, we designed an agentic evaluation on long horizon, end-to-end genomics analysis tasks. This evaluation assesses agentic performance on long-horizon quantitative tasks: based on realistic scientific data, can an agent plan valid analysis, QC, modeling, and corrections to arrive at decision-relative answers? Genebench spans a variety of domains, including functional genomics, spatial transcriptomics, proteomics, epigenomics, and applied genetics.

Table 2

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**
pass@1	21.6%	20.4%
Average output tokens	16721	24270

3.3 Applied life sciences research

In order to measure and continuously improve the real world impact of GPT-Rosalind, we designed **LifeSciBench**, an externally expert-judged benchmark on foundational aspects of life sciences research. Unlike existing benchmarks that evaluate a single component of model performance or biological domain in isolation, LifeSciBench takes an end-to-end view of scientifically-valuable tasks from practicing life sciences researchers. We use this benchmark to align progress with the needs and realities of beneficial life sciences research.

Table 3

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**
pass@1	63.4%	58.8%
Average output tokens	34439	33351

3.4 Labwork Bench

To test GPT-Rosalind-5.5's ability to help scientists conducting lab work in the real world, we designed an evaluation to measure model's ability to link perturbations to experimental outcomes in real wet lab protocols used by scientists, for the purposes ranging from troubleshooting to optimization. The data used by Labwork Bench is proprietary and thus uncontaminated. GPT-Rosalind-5.5 scores 63.2% vs. GPT-5.5 at 55.8%, while using fewer tokens.

Table 4

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**
pass@1	63.2%	55.8%
Average output tokens	45390.8	47950.8

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

4.1.1 Biological and Chemical

As we did for GPT-5.5, 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. A higher score on an evaluation indicates higher capability in that domain. As expected, GPT-Rosalind-5.5 scores meet or exceed GPT-5.5's capability of all the following evaluations but one (Multi-Select Virology Troubleshooting). We also show values for GPT-5.5 Pro, which is run with roughly an order of magnitude higher inference budget and thus is expected to exceed GPT-Rosalind-5.5 performance in some areas. In particular, GPT-Rosalind-5.5 was not specifically trained on virology or biothreat creation, and for these evaluations it expectedly underperforms GPT-5.5 Pro.

Table 5: **Overview of Biological and Chemical evaluations**

Evaluation	Capability	Description
ProtocolQA Open-Ended	Wet lab capabilities (open-ended)	How well can models perform on open-ended questions testing protocol troubleshooting?
TroubleshootingBench	Tacit knowledge and troubleshooting (open-ended)	Can models identify and fix real-world errors in expert-written lab protocols that rely on tacit knowledge?
Biorisk knowledge	Biorisk-relevant knowledge (closed-ended)	Can models answer closed-ended questions probing biosecurity-relevant biology and laboratory knowledge?
Multi-select virology troubleshooting	Wet lab troubleshooting (multi-select)	Can models identify likely causes and fixes for virology protocol failures when multiple correct options may apply?
Hard-negative protein binding prediction	Protein design	Can the model distinguish "gain of function" (binder) protein variants from non-functional (non-binder) proteins, given high quality 3D structure predictions and confidence scores from biological design tools?
DNA sequence design for transcription factor binding	DNA design	Can the model design DNA sequences to maximally bind known transcription factors, with better performance than freely available gradient-based methods?

4.1.1.1 ProtocolQA Open-Ended

Table 6: To evaluate models' ability to troubleshoot commonly published lab protocols, we modify 108 multiple choice questions from FutureHouse's ProtocolQA dataset [1] 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 for how to fix the procedure.

Metric	GPT-Rosalind-5.5	GPT-5.5	GPT-5.5 Pro
pass@1	36.4%	37.3%	36.73%

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 (virology, genetics, microbiology, or protein engineering) were asked to transcribe biological protocols they have personally used in the lab. Each protocol must include precise step-by-step procedures, equipment, and reagents. If a protocol was adapted from a publication, experts were required to significantly alter at least several steps. From these protocols, they created three troubleshooting questions each,

introducing subtle or realistic execution errors (e.g., improper homogenization technique) and describing the resulting failed outcome.

After going through independent expert review, the resulting dataset includes 52 protocols, each paired with three expert-written troubleshooting questions.

Table 7: Compared to ProtocolQA Open-Ended, which focuses on well-known published procedures, TroubleshootingBench is designed to test model performance on non-public, experience-grounded protocols and errors that rely on tacit procedural knowledge

Metric	GPT-Rosalind-5.5	GPT-5.5	GPT-5.5 Pro
pass@1	53.31%	44.1%	49.58%

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 5 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, i.e., they either require tracking down authors of relevant papers or knowing people in the field. Troubleshooting questions are meant to be obscure to anyone without hands-on experience, i.e., the answers are known only to people who have tried the protocol themselves.

Table 8: This set is uncontaminated; it was created fully in-house with our partners at Gryphon Scientific and has not been published.

Metric	GPT-Rosalind-5.5	GPT-5.5	GPT-5.5 Pro
cons@32	81.1%	78.3%	81.67%

4.1.1.4 Multi-select Virology Troubleshooting

Table 9: 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.

Metric	GPT-Rosalind-5.5	GPT-5.5	GPT-5.5 Pro
pass@1	51.2%	53.7%	55.34%

4.1.1.5 Hard negative protein binding prediction

Offensive capabilities in an AI system likely require expert sequence-to-function knowledge, which can be measured by its ability to distinguish gain-of-function sequences using information beyond that provided by sophisticated biological design tools. We probe this capability using a proprietary, non-contaminated dataset of protein-binding experimental results spanning 43 unique protein targets and 492 hotspots. In this task, gain of function is operationalized as binding capability: the model must distinguish functional (binding) proteins from non-functional (non-binding) proteins.

We structure each problem to minimize the effectiveness of trivial data-science heuristics and thereby measure uplift attributable to expert biochemical knowledge. In each task, the model is given five candidate binders designed against the same hotspot on a target protein, with pairwise sequence similarity of at least 0.85. All candidates in this setting had AlphaFold 2 interface predicted TM (ipTM) scores of at least 0.8, indicating high-confidence, high-quality binding predictions. This setup removes the trivial strategy of selecting candidates solely on the basis of a high ipTM score.

The model was additionally given 3D structures of the predicted protein-protein binding complexes as PDB files, along with access to a computer and a browser to fully elicit its tool-use and reasoning capabilities in a long-horizon, agentic setting. The model is not told how many true binders exist among the set of 5, while the true number was held at 1 to properly reflect the rarity of true binders and to avoid uplift from simple heuristics such as clustering.

Table 10: Biological experiments are costly and iterative. Because errors compound across steps, a single low-success-rate step can substantially constrain a project's overall probability of success. We hypothesize that a qualitative capability shift is most likely to emerge when stepwise success rates exceed a relatively high threshold. Accordingly, we propose 50% correctness as the threshold for biorisk concern.

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**	**GPT-5.5 Pro**
pass@4	3.13%	0.4%	0.0%

4.1.1.6 DNA sequence design for Transcription Factor binding

We evaluate the model's ability to design DNA sequences with high binding affinity for well-known transcription factors (TFs), benchmarking it against widely-available gradient-based design methods. Designing sequences with high binding affinity to TFs could allow modulation of gene expression through gene-editing approaches. We evaluate the model on 11 TFs drawn from Nucleobench, creating 50 prompts per TF. Each prompt contains a starting sequence of 3000 basepairs chosen at random from an {A,C,G,T} vocabulary. Generated sequences are scored using high-performance oracles from the TF-specific models in the BPNet family, with Basenji2 models as secondary oracles when available for the TF of interest.

We benchmark model performance against a freely available and simple gradient-based approach, Ledidi, which was found to be competitive in Nucleobench for many sequence-design tasks. We set a threshold of 80% win rate over Ledidi for significant DNA design capabilities. We find that GPT-Rosalind-5.5 performs significantly below this baseline.

Table 11

Metric	**GPT-Rosalind-5.5**	**GPT-5.5**	**GPT-5.5 Pro**
pass@1	13.64%	13.82 %	16.5%

4.2 Safeguards

Compared to GPT-5.5, GPT-Rosalind-5.5 features additional deployment controls for its use in advanced life sciences research. GPT-Rosalind-5.5 is a limited, controlled deployment; 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 controlled access with enterprise-grade security.

Our safeguard approach for GPT-Rosalind-5.5 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. Unlike our safeguards posture for our more broadly distributed flagship models, in this case we are not deploying automated monitors for real-time blocking of potentially unsafe generations.

4.2.1 Trust-based access

GPT-Rosalind-5.5 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.

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 has also introduced a trusted-access pathway for select government entities working on approved biosecurity, biosafety, public-health preparedness, and biodefense use cases, as described in a recent blog post. This pathway is intended to help responsible government users apply advanced life sciences capabilities to preparedness-oriented work, such as improving public-health readiness, strengthening biosafety and biosecurity, supporting biological threat assessment, and accelerating beneficial medical countermeasure or response-related research.

Approved customers must limit GPT-Rosalind-5.5 access to authorized users with a legitimate need and apply least-privilege access controls, appropriate onboarding, and insider risk controls to their users.

4.2.2 Model-level boundaries against harmful biological and cyber assistance

4.2.2.1 Biological and Chemical boundaries

GPT-Rosalind-5.5 is designed to be more useful for legitimate advanced biological research than generally available ChatGPT models. This includes support for advanced life sciences workflows such as literature and knowledge retrieval, data parsing and analysis, computational biology, and use of specialized biology tools and databases.

Like GPT-5.5, GPT-Rosalind-5.5 is trained to refuse malicious requests that would meaningfully enable biological weaponization.

Because some advanced biological capabilities are dual-use, GPT-Rosalind-5.5 is not broadly available. Access is gated through the trusted access review process described above.

4.2.2.2 Cyber boundaries

GPT-Rosalind-5.5 inherits the model-level cyber safeguards of the GPT-5.5 base model.

The model continues 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.

OpenAI will monitor for cyber misuse and investigate activity that appears misaligned with the customer's approved life sciences use case. Confirmed malicious cyber activity may result in enforcement under OpenAI's Usage Policies, including potential account banning or revocation of GPT-Rosalind-5.5 access.

4.2.3 Monitoring, Review, and Revocation

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 biological weaponization, malicious cyber activity, unauthorized access, or use that is materially inconsistent with the customer's approved purpose.

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.

Because advanced life sciences work is inherently dual-use, some legitimate research activity may require contextual review. Our process considers both the risk of the activity and the apparent purpose of the use, including whether it remains aligned with the customer's approved life sciences, public-health, biosafety, biosecurity, or biodefense use case. Activity that appears to cross into biological weaponization, malicious cyber activity, unauthorized access, or other prohibited harmful use may be flagged for review and action. OpenAI also seeks to monitor for unresolved customer security concerns, failure to maintain required access controls, or material changes in approved use case, deployment pathway, or user population.

OpenAI reserves the option to apply additional monitoring, request information from the customer, narrow or restrict access, require remediation of security or governance issues, or revoke GPT-Rosalind-5.5 access.

4.2.4 Security and Access Controls

Approved GPT-Rosalind-5.5 customers must maintain security controls appropriate for access to a higher-capability life sciences model. These include strong identity and access management, administrator controls, role-based access controls, and processes to remove access when a user no longer has a legitimate need.

These controls are intended to help ensure that GPT-Rosalind-5.5 is used only by approved users, for approved purposes, within accountable organizations.

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-5.5 safeguards as the deployment evolves.

References

[1] 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.