BioSecBench-RespiratoryVirus-Surveillance

Benchmarking AI agents on the analyses used to detect, characterize, and track respiratory viruses

Harmon Bhasin[1], Evan Seeyave[1], Gage Moreno[1], Arjun Banerjee[1], Ioanna Pagani[1], Tyler Chafin[1], Paul Oluniyi[1], Pedro Belda-Ferre[1], Lucas Yago Melo Ferreira[1], Zack Mudge[1], Marcos Parras Molto[1], Tristan Russell[1], Karl Widney[1], Mikail Kelleci[1], Konstantinos Konstantinidis[1], Reema Singh[1], Naomi Mora Jaramillo[1], Sara Zufan[1], Michael Saldana[1], Arya Kaul[1], Dianzhuo Wang[1], Kenny Workman[1]

[1]LatchBio, San Francisco, CA, USA

Correspondence: harmon@latch.bio, kenny@latch.bio

Abstract

Respiratory viruses cause recurrent epidemics and can trigger pandemics, making reliable detection and surveillance a continuing public-health priority. AI agents could support the diagnostic, genomic, and epidemiological analyses underlying this work, but their reliability across these tasks remains uncertain. We introduce BioSecBench-RespiratoryVirus-Surveillance, a benchmark of 110 evaluations examining how respiratory viruses are detected, characterized, and tracked across populations. Each evaluation pairs a surveillance question and source materials with task-specific grading criteria for numerical, categorical or structured outputs. Across 34 model–harness configurations (21 models, three offline runs per evaluation), the highest preliminary overall pass rate was 25.5% (Opus 5.5 / PI), followed by Opus 5.5 / Claude Code (19.4%) and Gemini 4 Argon / PI (17.6%). Six models led or tied for the lead in at least one of the 17 displayed evidence-source, decision-type or failure-class groups, and the highest mean pass rate among the six challenge classes was 19.2%.

Introduction

Respiratory viruses cause recurrent epidemics and can trigger pandemics [1], creating a continuing need to detect emerging threats and track ongoing disease burden [1, 2, 3]. Surveillance and epidemiology inform public-health responses by establishing which viruses are circulating, how they are spreading, and whom they affect [1, 4]. This work requires analyses across laboratory, genomic, and population-level evidence. Diagnostic scientists assess whether assays reliably detect circulating viruses, including when sequence variation affects assay targets [5, 6]. Bioinformaticians use sequencing data to identify pathogens [7], characterize viral variation [8], and investigate transmission [9], while epidemiologists use clinical and environmental observations to estimate circulation and burden [4, 10] and forecast future activity [11, 12].

AI agents that combine planning, specialized tools, and code execution could support these diagnostic, genomic, and epidemiological analyses. Existing benchmarks evaluate biological data analysis, bioinformatics workflows, and research-study-scale tasks [13, 14], while BioSecBench-Surveillance evaluates agents on analysis pipelines for pathogen genomic surveillance [15]. These efforts motivate a more focused question: how reliably can agents perform the range of analyses needed for respiratory-virus surveillance and epidemiology? We introduce BioSecBench-RespiratoryVirus-Surveillance, a benchmark of 110 evaluations spanning diagnostic assay development and evaluation; pathogen discovery and characterization; viral evolution and transmission; and population circulation, disease burden, and forecasting. Each evaluation pairs a scientific question and source materials with task-specific grading criteria. We assess performance across scientific categories to understand how reliably agents perform these analyses and where their limitations could affect efforts to detect emerging respiratory threats and track disease burden.

Benchmark Construction

The benchmark contains 110 evaluations across four areas of respiratory-virus surveillance and epidemiology: diagnostic assay development and evaluation (25), pathogen discovery and microbial-community characterization (25), viral evolution and transmission (30), and population circulation, disease burden, and forecasting (30). Tasks concern SARS-CoV-2, respiratory syncytial virus, influenza, and other respiratory viruses, drawing on genomic sequences, diagnostic measurements, clinical records, and environmental and animal monitoring data. Figure 2 summarizes the cohort by scientific category, viral surveillance targets, evidence source, and requested decision.

Each evaluation asks agents to answer a scientific question using a published dataset available online. We tested 34 model–harness configurations spanning 21 models, with three offline runs per evaluation using PI, Claude Code, Codex, or Grok Build. We graded the requested outputs using task-specific numerical tolerances, categorical matching, and structured-output checks. We report pass rates overall and by scientific category and failure class.

Results

Leaderboard and efficiency

The highest overall pass rate was 25.5%, achieved by Opus 5.5 / PI (95% CI 18.8–32.1; Figure 1A). Opus 5.5 / Claude Code followed at 19.4%, then Gemini 4 Argon / PI at 17.6%. The two leading configurations used the same model, but its pass rate was 6.1 percentage points higher with PI than with Claude Code.

Pass rates were similar between harnesses for most models, while larger differences favored different harnesses depending on the model. Seven of the 13 models tested under two harnesses had pass rates within one percentage point. GPT-6.1 Terra scored 4.5 percentage points higher with Codex than with PI, whereas GPT-6 Astra scored 3.0 points higher with PI than with Codex.

Similar performance also came at different resource costs (Figure 1B–D). Gemini 3.8 Flash / PI and Nemotron 3 Ultra 550B A55B / PI both achieved a 7.3% pass rate, but Gemini averaged $1.24 per run compared with $0.22 for Nemotron. Gemini also generated 56.0 thousand output tokens per run compared with 12.8 thousand, and took 31.3 minutes compared with 10.4 minutes. In this comparison, greater resource use did not yield a higher pass rate.

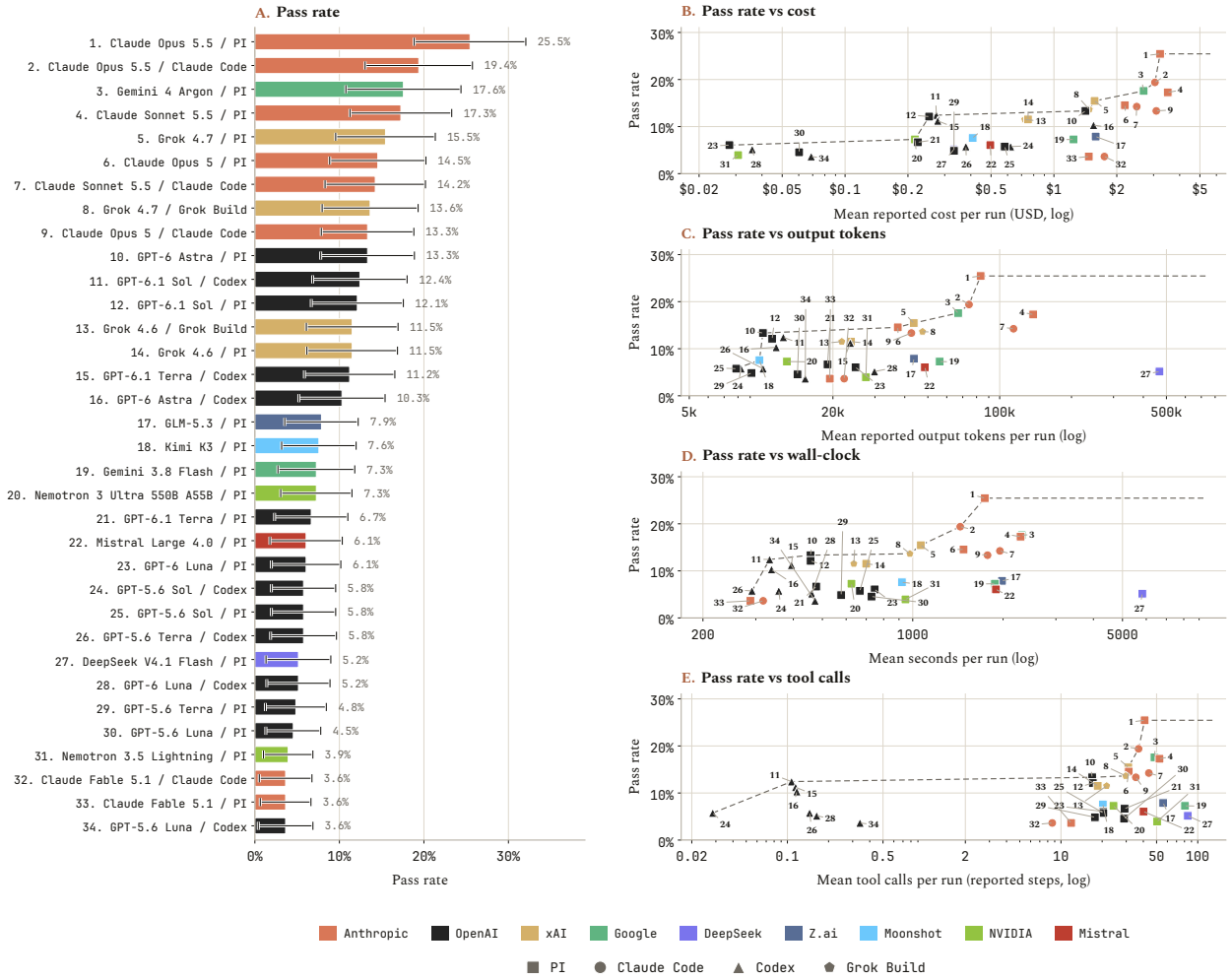


Figure 1: Pass rate by configuration. (A) Mean pass rate over the 110 evaluations, with 95% Student's t intervals taken over per-evaluation pass rates; colour marks the model's lab. Refusals count as failures. (B–E) Pass rate against mean reported cost, output tokens, wall-clock time and tool calls per completed, non-refused run, on log scales. The dashed line marks the efficiency frontier.

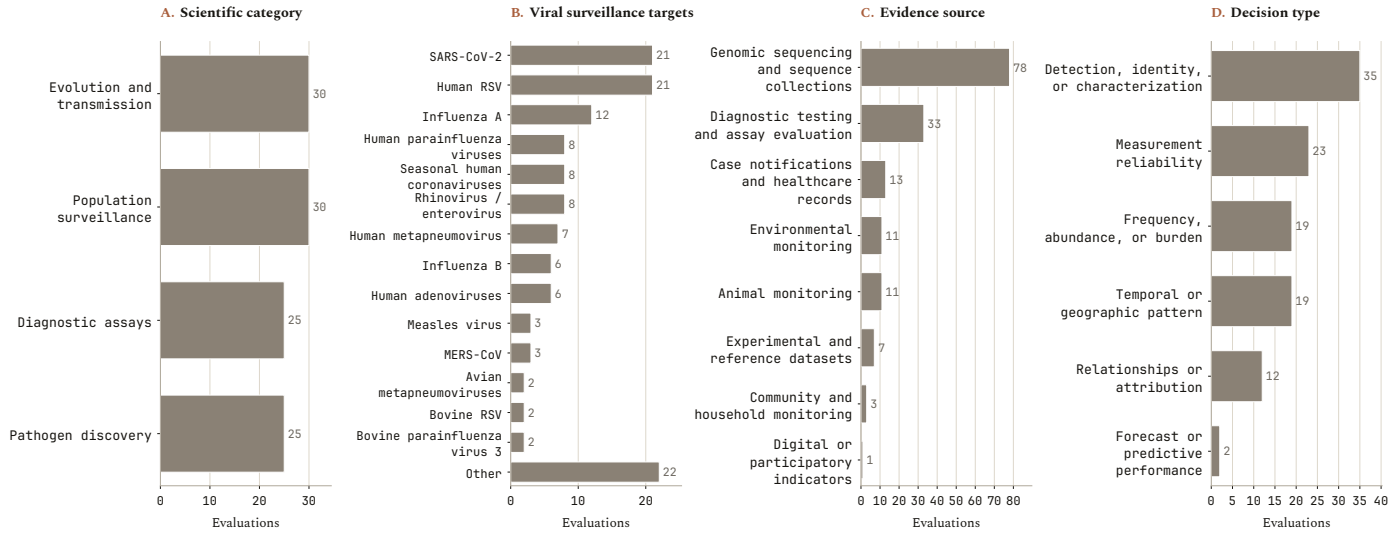


Figure 2: Benchmark anatomy. All 110 selected evaluations by (A) scientific category, the area of surveillance analysis; (B) viral surveillance targets, named viruses or defined surveillance groups; (C) evidence source, the types of observations or measurements analyzed; and (D) decision type, the requested analytical endpoint. Categories and decision types have one label per evaluation. Targets and evidence sources permit multiple labels, with each evaluation counted once per applicable label. Other includes targets represented in only one evaluation and unspecified respiratory viral targets.

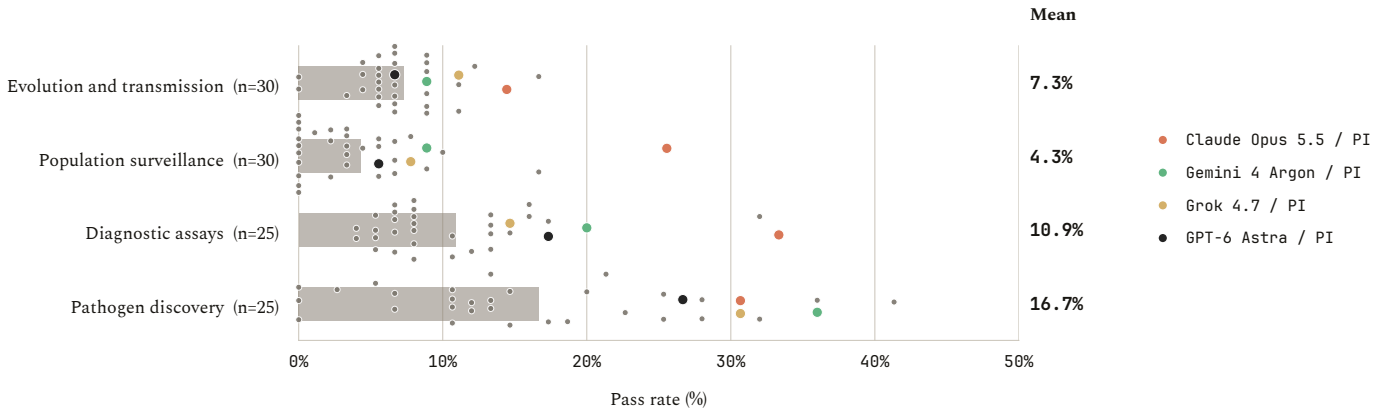


Figure 3: Pass rate by scientific category. Bars show the mean across 21 models, pooling available harness counts within each model and evaluation before averaging evaluation pass rates. Dots show individual model–harness configurations. Colour highlights the best configuration from each of the top four labs, ranked by overall pass rate; other configurations are grey. Category labels give evaluation counts. Refusals count as failures.

Performance by category

Mean pass rates were highest for pathogen discovery and lowest for population surveillance (Figure 3). Averaged across the 21 models, pass rates were 16.7% for pathogen discovery, 10.9% for diagnostic assays, 7.3% for evolution and transmission, and 4.3% for population surveillance. This pattern held across most models: discovery was the highest-scoring category for 16 of 21 models, including one tie, while population surveillance was the lowest for 16, including two ties.

Opus 5.5 had the highest overall pass rate, but other models led on specific evidence sources, decision types, and failure classes (Figure 4). Six models led or tied for the lead in at least one of the 17 displayed groups. Grok 4.7 scored 34.8% on animal-monitoring tasks, compared with 15.2% for Opus 5.5. Sonnet 5.5 led on evaluations grouped under unsupported identity or transmission claims at 38.9%, compared with 19.4% for Opus 5.5.

Healthcare-record analyses and temporal or geographic decisions had low mean pass rates across models. For evaluations using case notifications and healthcare records, the mean pass rate was 1.8% and the highest model score was 7.7%. For evaluations asking models to identify temporal or geographic patterns, the mean pass rate was 2.0%, and 13 of the 21 models passed none of the 19 evaluations.

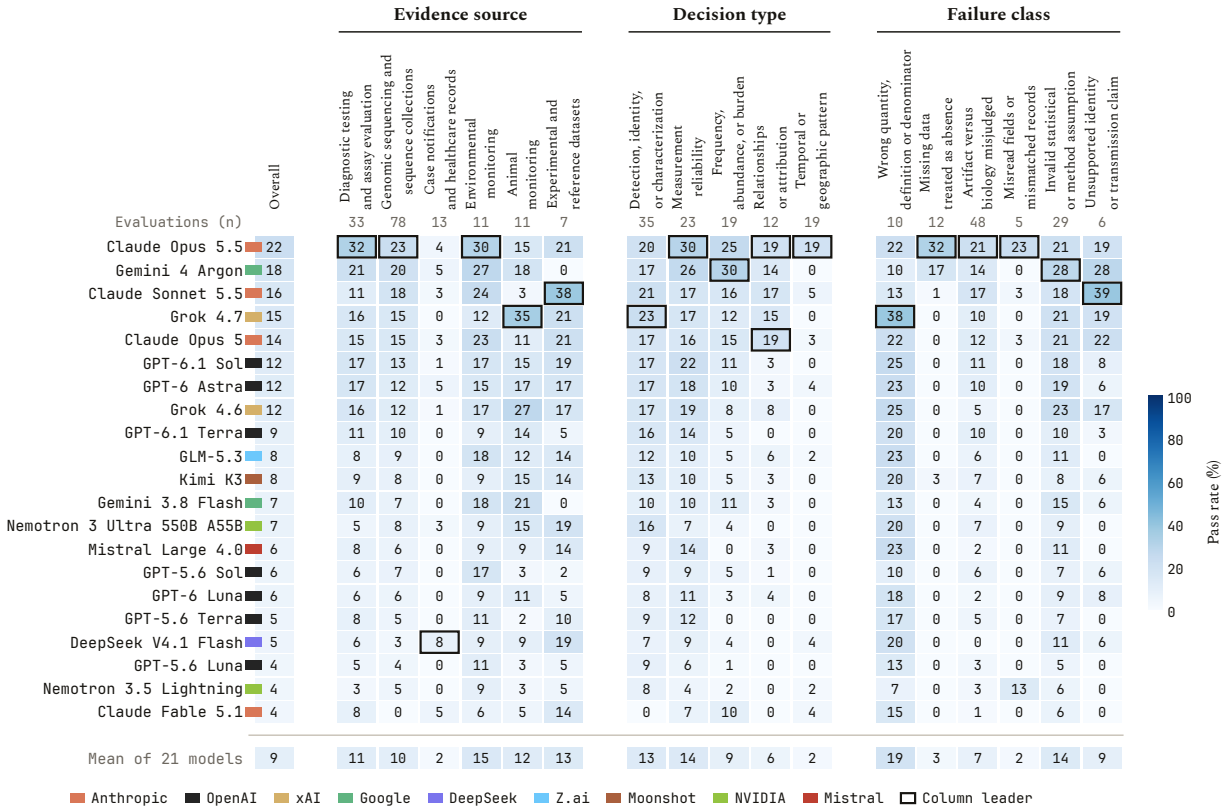


Figure 4: Pass rate by evidence source, decision type and failure class. All 21 models, with available harnesses pooled within each evaluation, ordered by overall pass rate. Evidence sources can overlap; decision type and failure class each assign one label per evaluation. Outlined cells mark column leaders, including ties, and the bottom row is the mean over models. Columns with fewer than five selected evaluations are omitted and cells with fewer than five observed evaluations are blank. Refusals count as failures.

Failure decomposition

Mean pass rates were below 20% in every failure class (Figure 5). Performance was lowest on evaluations grouped under misread fields or mismatched records (2.1%) and missing data treated as absence (2.5%). In each group, 17 of 21 models passed no evaluations (Figure 4). By comparison, mean pass rates were 19.0% for wrong quantities or denominators and 13.6% for invalid statistical or method assumptions. Every model passed at least one evaluation in each of these two groups.

Distinguishing artifacts from biological signals also remained difficult across models. This was the largest observed failure class, covering 48 of 110 evaluations, with a mean pass rate of 7.5%. It included errors in both directions: accepting artifacts as biological signals and dismissing genuine signals as artifacts. Opus 5.5 led the group at 21.2%, indicating that the low average was not solely attributable to weaker models.

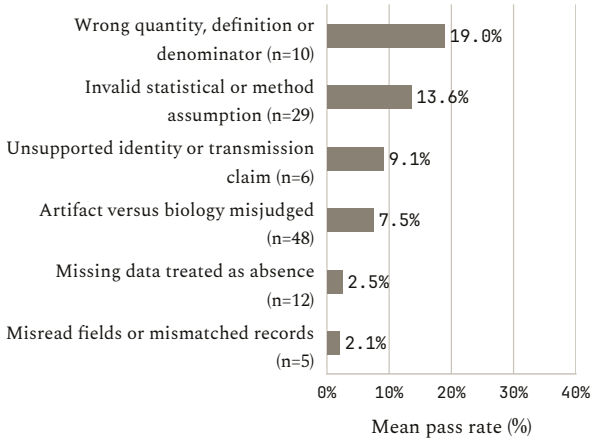


Figure 5: Mean pass rate by failure class. Mean over the 21 models of each model's pass rate on the evaluations in the class, with available harnesses pooled within each evaluation; n is the number of evaluations in the class. Refusals count as failures.

Discussion

Current agents do not yet reliably solve these respiratory-virus surveillance tasks. The highest overall pass rate was 25.5%. This measure counts refusals as failures to reflect whether an agent completes the requested analysis, so models that refuse more often receive lower scores. Performance also varied across scientific categories, with population surveillance showing the lowest mean pass rate across models. Review of failed executions helps characterize the scientific difficulties behind these results. Mean pass rates were lowest on evaluations where reviewed failures most often involved misreading fields, mismatching records, or treating missing observations as negative results. Distinguishing artifacts from biological signals was the largest observed failure class, assigned to 48 of 110 evaluations based on their reviewed failed executions. Together, these findings identify interpretation of the underlying records and measurements as a recurring limitation in agents' surveillance analyses.

A main limitation is the uneven coverage of respiratory viruses and surveillance analyses. Coverage is shaped by the availability of suitable published datasets, with some viral targets, evidence sources, and types of analysis represented more frequently than others. For example, genomic evidence is prominent, while forecasting and community-based surveillance receive less coverage. This limits how broadly the results can be generalized across surveillance settings. Future expansion should prioritize these gaps to provide a more complete assessment of agent capabilities across respiratory-virus surveillance.

We believe BioSecBench-RespiratoryVirus-Surveillance will be a useful tool for practitioners assessing AI agents for respiratory-virus surveillance and epidemiology. Its combination of task-level evaluation and failure analysis provides a basis for comparing capabilities relevant to their work. As agents improve, the benchmark can help practitioners assess whether those improvements translate into more reliable respiratory-virus surveillance.

References

1. World Health Organization. "Crafting the mosaic": a framework for resilient surveillance for respiratory viruses of epidemic and pandemic potential. Geneva: World Health Organization, 2023.
2. Iuliano AD, Roguski KM, Chang HH, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*. 2018;391:1285–1300.
3. Li Y, Wang X, Blau DM, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. *The Lancet*. 2022;399:2047–2064.
4. European Centre for Disease Prevention and Control. Integrated respiratory virus surveillance guidance for the EU/EEA: A framework for surveillance to inform public health action. Version 1.0. Stockholm: ECDC, 2026.
5. Leber AL, Everhart K, Daly JA, et al. Multicenter Evaluation of BioFire FilmArray Respiratory Panel 2 for Detection of Viruses and Bacteria in Nasopharyngeal Swab Samples. *Journal of Clinical Microbiology*. 2018;56:e01945-17.
6. Ziegler K, Steininger P, Ziegler R, et al. SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene. *Eurosurveillance*. 2020;25:2001650.
7. Langelier C, Kalantar KL, Moazed F, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proceedings of the National Academy of Sciences*. 2018;115:E12353–E12362.
8. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34:4121–4123.
9. du Plessis L, McCrone JT, Zarebski AE, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*. 2021;371:708–712.
10. Boehm AB, Hughes B, Duong D, et al. Wastewater concentrations of human influenza, metapneumovirus, parainfluenza, respiratory syncytial virus, rhinovirus, and seasonal coronavirus nucleic-acids during the COVID-19 pandemic: a surveillance study. *The Lancet Microbe*. 2023;4:e340–e348.
11. Reich NG, McGowan CJ, Yamana TK, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology*. 2019;15:e1007486.
12. Cramer EY, Ray EL, Lopez VK, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*. 2022;119:e2113561119.
13. Zou C, Banerjee A, Bhasin H, et al. OmicsBench: A Large-Scale, Verifiable Benchmark for Agentic Life Sciences Capabilities. LatchBio, 2026. Technical report.
14. Koch Z, Wassie AT, Valdes-Aleman J, et al. BixBench3: Benchmarking AI agents on research-study-scale computational biology tasks. arXiv:2608.25286, 2026. Preprint.
15. Bhasin H, Flyangolts K, Wang D, et al. BioSecBench-Surveillance: A Verifiable Benchmark for AI Agents in Pathogen Genomic Surveillance. arXiv:2607.19262, 2026. Preprint.