ICML 2026

BioAgent Bench

An AI Agent Evaluation Suite for Bioinformatics

{ } github.com/bioagent-bench/bioagent-bench

[PDF] arXiv:2601.21800

Dionizije Fa

Marko Čuljak

Bruno Pandža

Mateo Čupić

ICML 2026

From bioinformatics Q&A to robust workflow execution

MOST PRIOR BENCHMARKS ASK

"Can the model answer biology questions?"

question → answer

BIOAGENT BENCH ASKS

"Can an agent execute a multi-step bioinformatics pipeline, produce the right artifacts and outputs, and remain robust when conditions change?"

plan	tool calls	files
artifacts	outputs	trace

ICML 2026

BioAgent Bench

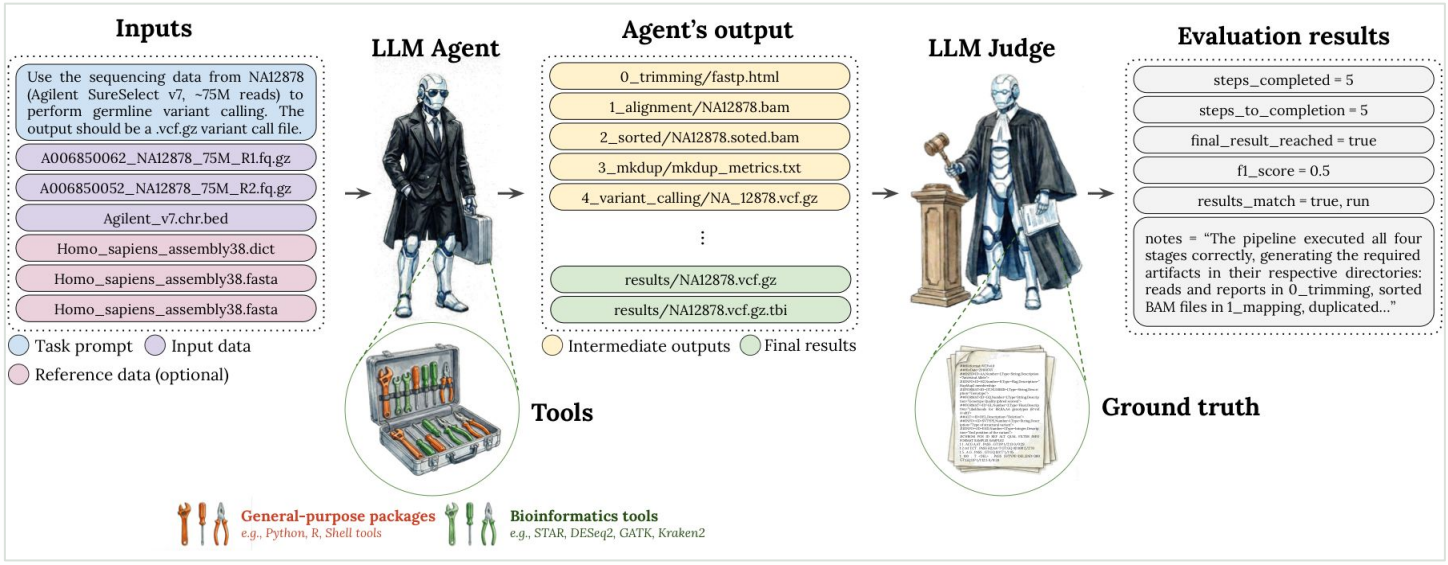
TASK SUITE

10 end-to-end bioinformatics tasks

Evaluation checks artifacts, outputs, traces, stability under perturbation

- **alzheimer-mouse** — *Mus musculus*
- **transcript-quant** — *Homo sapiens (simulated)*
- **comparative-genomics** — *Micrococcus spp.*
- **viral-metagenomics** — *Viruses from Tursiops truncatus*
- **cystic-fibrosis** — *Homo sapiens*
- **deseq** — *Candida parapsilosis*
- **evolution** — *Escherichia coli*
- **giab** — *Homo sapiens*
- **metagenomics** — *Cuatro Cinenegas Basin*
- **single-cell** — *Homo sapiens*

EVALUATION PIPELINE



Perturbation testing

AFTER GRADING

- INPUT PERTURBATIONS
- TOOL / ENV PERTURBATIONS
- SCORING ROBUSTNESS

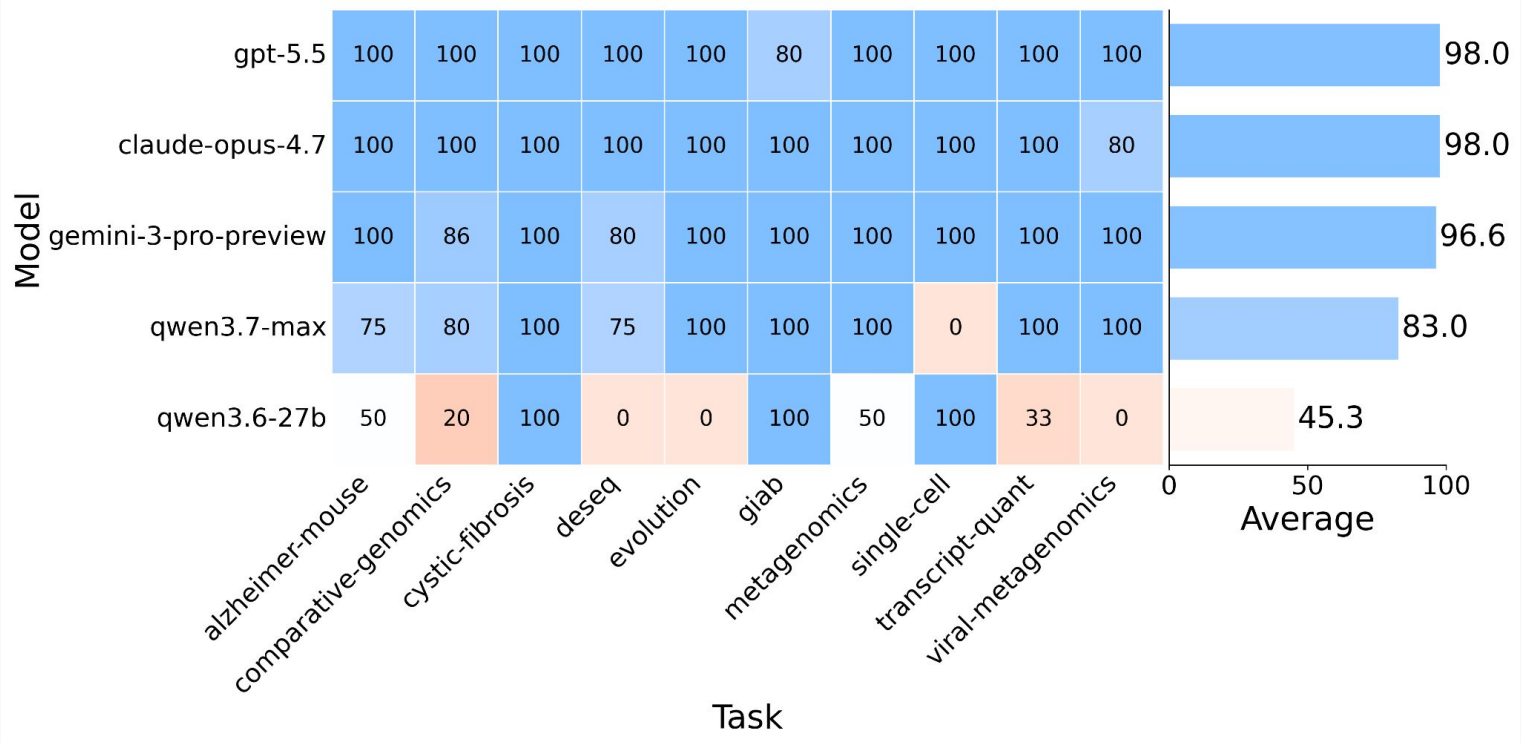
Agents can complete real pipelines

PRIMARY METRIC

completion rate

Completion = required pipeline steps + final output artifact passing the grader.

Frontier agents mostly reach the requested final artifact without custom scaffolding.



ICML 2026

Completion drops when conditions change

We've answered "can it finish?" Perturbation testing asks whether it can finish reliably, for the right reasons.

Robustness across repeated trials

Task	Trials	Jaccard	Numeric overlap
alzheimer	4	0.160	0.219
comparative	4	0.004	NA
cystic-fibrosis	3	1.000	NA
deseq	4	0.978	0.995
evolution	4	0.000	NA
metagenomics	4	0.395	0.746
single-cell	4	0.114	0.395
transcript-quant	4	1.000	1.000
viral-metagenomics	4	0.667	1.000

Stress tests expose jagged intelligence

↓ LOWER IS BETTER

Prompt bloat

-28%

completion

Longer prompts reduce average completion by 28%.

Corrupt inputs

3/10

went undetected

Corrupt files were flagged in most tasks, but not all.

Decoy files

2/10

used

Agents still incorporate wrong files into the final workflow.

ICML 2026

Main Takeaways

READINESS METRIC

- Completion alone is necessary but insufficient for real-world readiness.
- Agents can construct pipelines and final artifacts while still missing step-level reasoning failures, such as incorrect file selection or ignored corrupted inputs.
- In clinical or regulated settings, agents must justify choices and avoid proceeding when evidence is unreliable.

LIMITATIONS

Grader design: how do we verify unverifiable biology?