RareDiseaseBench

Can agents request, analyze, and interpret data to diagnose rare disease patients?

You Chen[1], Arjun Banerjee[1], Yuyi Liu[1], Parnal Joshi[1], Narendrareddy Munagala[1], Martin Jacko[1], Christopher Zou[1]

[1]LatchBio, San Francisco, CA, USA

Correspondence: chris.zou@latch.bio

Abstract

Rare diseases cumulatively affect 300 million individuals worldwide but are difficult to diagnose, creating burden for patients, families, and communities. AI agents show promise in automating and scaling one bottleneck to diagnosis: the analysis of complex biomedical data. We present LatchBio's first RareDiseaseBench, a benchmark of 35 rubric-graded evaluations in which agents must integrate clinical, multiomics, and functional assay data from patients to draft diagnosis and management plans over one or more visits. For better clinical and scientific realism, RareDiseaseBench tasks run in a system where agents are not provided with all of the relevant data upfront; they must request the right data to analyze. Across 31 model–harness configurations (20 models, three offline runs per evaluation), the highest preliminary overall pass rate was 43.8% (Gemini 4 Argon / PI), followed by Claude Opus 5.5 / Claude Code (33.3%) and GPT-6 Astra / PI (33.3%). Performance varies across genetic architectures, variant types, and affected organ systems. RareDiseaseBench provides a standard for measuring agentic performance in a domain where rapid AI adoption could provide huge benefit for researchers, clinicians, and patients.

Introduction

300 million individuals worldwide are affected by 7,000 identified rare diseases. Despite their cumulative prevalence, rare diseases are notoriously difficult to diagnose [1]. By definition, they are locally rare (affecting fewer than 1 in 2,000 people in each WHO region). They are also clinically heterogeneous and understood in depth by fewer physicians [2]. As a result, even well-resourced rare disease patients often experience a long diagnostic odyssey of referrals, misdiagnoses, and unnecessary treatments [3,4].

Rapid improvements to sequencing costs, accuracy, and scalability have made multiomics and functional assays increasingly available to patients and clinicians. These data, especially various forms of whole-genome sequencing, have been found in large-scale cohort studies to dramatically improve rare disease diagnostic yield [5]. However, the volume and complexity of data generated poses new challenges in conducting accurate and interpretable bioinformatics analysis at scale [6].

Our previous work has examined strengths and weaknesses of agents powered by large language models (LLMs) in analyzing and reasoning about data across bioinformatics, therapeutics development, and oncology [7–11]. Others have examined whether agents can conduct clinical practice [12], improve diagnostic yield for unsolved rare disease cases [13], analyze multimodal rare disease cases [14] and find causal variants in patient genomes [15]. However, these benchmarks have a shared design choice: they all provide a task prompt and collection of data objects upfront to an agent. This design choice has two consequences. First, the composition of the data collection leaks information about the task. Second, it is impossible to examine whether agents could have sourced the information required to solve the task. In real-world settings, identifying the data required to make a diagnosis while under feasibility constraints is a critical capability.

Here, we introduce the first edition of RareDiseaseBench, a collection of 35 single- and multistep rare disease diagnosis tasks that require an agent to request, analyze, and integrate clinical, multiomics, and functional data to write diagnoses and management plans. Each task is based on one or more previously published patient studies and graded with a task-specific rubric. Across the benchmark, tasks span cases with different genetic architectures and data modalities. We assess performance across segments of the benchmark and provide case studies useful for understanding LLM utility in rare disease diagnosis.

Benchmark Design

Each RareDiseaseBench task asks agents to develop a diagnosis and management plan given a task prompt and upfront data objects (Figure 1A). Prompts are organized into one or more steps akin to Harbor multi-step [16]. Tasks also specify contents of an isolated Records system, which includes BAM files, images, markdown reports, and other artifacts necessary to solve the case, along with natural language descriptions of conditions for their release. To access these files, agents are provided with a `get_info` tool that takes in information and justification arguments. `get_info` calls are routed to a service with visibility into the Records system. Upon receiving a request, the service first uses an LLM judge (Sonnet 5.5) to determine eligible files based on the justification. Next, it checks whether the requested files are eligible based on previous requests: for example, the agent cannot ask for immunohistochemistry stains on a biopsy sample before it has asked for a biopsy. If both checks pass, the files are copied from the Records system into the agent's workspace. After submitting an answer, agents' diagnosis and management plans are graded by vote among an ensemble of 3 frontier LLM judges (Gemini 3.8, Sonnet 5.5, and GPT-6.1-Terra, medium reasoning) against an expert-developed rubric. Figure 1A shows the RareDiseaseBench harness.

The 35 RareDiseaseBench tasks are drawn from previously published cases [17–51] and constructed using synthetic genomes except in one case where locus-targeted reads were publicly available. Tasks vary across genetic architectures, variant types, affected organ systems, and data modalities. Of the 35 tasks, 30 include exome sequencing, 29 short-read genome sequencing, 6 long-read sequencing, 3 RNA sequencing, 20 functional readouts reported as text, 5 methylation data, and 7 clinical images. Figure 1B–E summarizes the benchmark by genetic architecture, variant type, affected organ system, and inheritance mode.

Before inclusion in the benchmark, tasks underwent checks for rubric and service judge consistency: rubrics were required to be 100% consistent between all three LLM judges across 10 judgements on validation outputs, while the Records system had to pass at least three test requests for every data object: a successful release, a denied release, and an adversarial release where the requesting agent attempted to game the Records system judge. All tasks in the benchmark underwent review by experts holding a PhD or MD with expertise in the field. Here, we present results from running RareDiseaseBench on 31 model–harness configurations spanning 20 models, with three offline

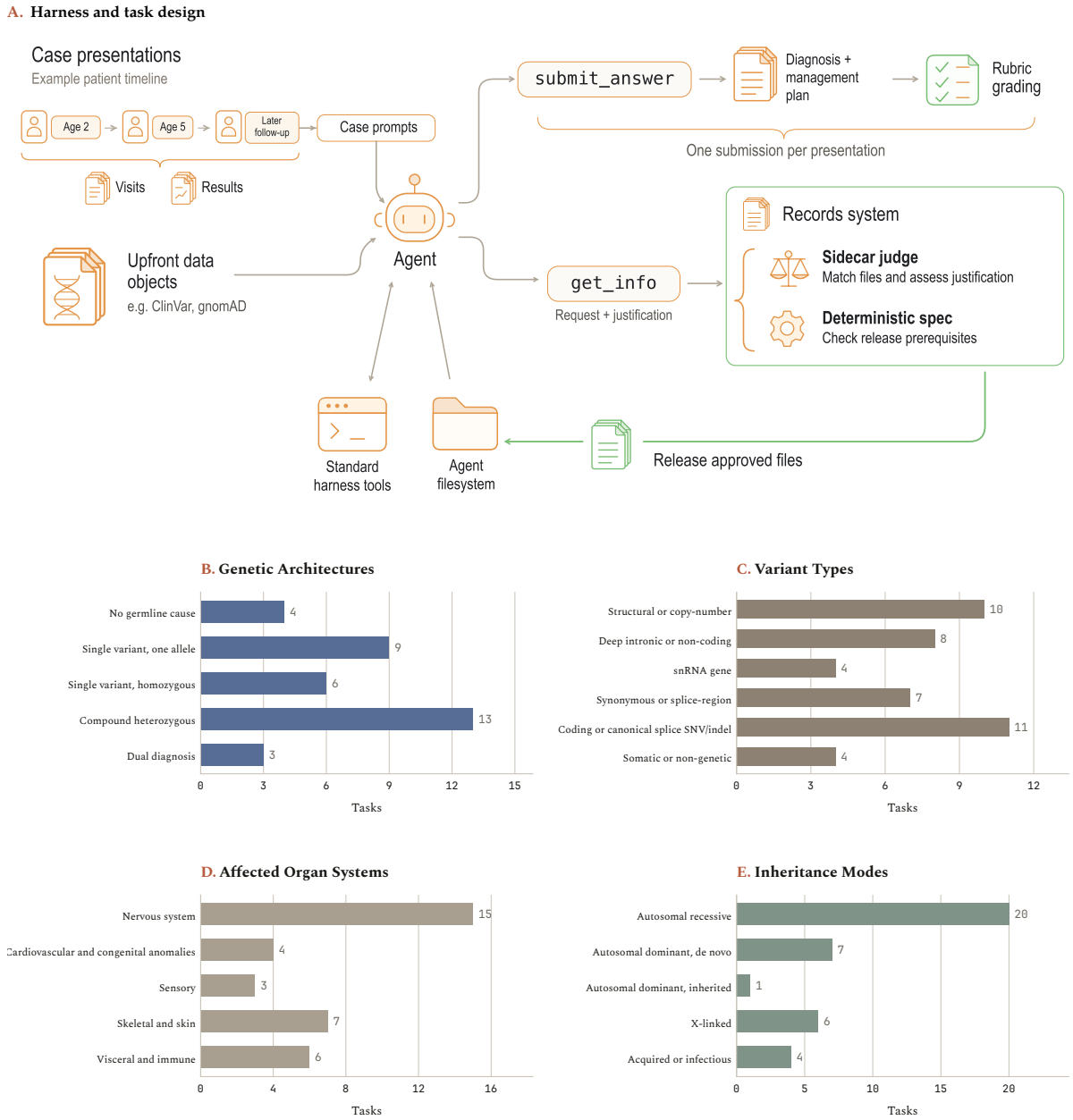


Figure 1: RareDiseaseBench harness, task design and anatomy. (A) Agents received a series of prompts and upfront reference data objects. In addition to the standard filesystem and coding tools provided by existing harnesses, we provided `get_info` and `submit_answer` tools. Agents could use `get_info` to request genomic and functional assay outputs from an isolated records system. A sidecar judge and deterministic filters would either accept or deny the request, placing the appropriate files in the agent's filesystem. Diagnosis and management plans were graded based on a per-task rubric. (B–E) Number of tasks per genetic architecture type, variant type, affected organ system and inheritance mode. Each task has one genetic architecture and one organ system; a task may carry several variant types and, for a dual diagnosis, several inheritance modes.

runs per evaluation. All Claude runs were executed through the LSVP program. Runs were limited to 6 hours; a run that did not finish within its limit counts as a failure. We report pass rates—where a pass indicates passing every rubric item and timeouts are considered failures—examine failure modes, and detail several case studies.

Results

Leaderboard and efficiency

The highest pass rate was 43.8% (Gemini 4 Argon / PI) (Figure 2A). Opus 5.5 / Claude Code and GPT-6 Astra / PI followed at 33.3%. 9 of 31 configurations passed no attempt at all, including all but one configuration based on an open-source model. 29 of the 35 tasks were solved by at least one configuration, 26 by models from two or more labs, and 16 were passed in all three attempts by some configuration.

Model-specific harnesses produced performance gains in Claude models, but not GPT-series models. Claude models achieved higher or equivalent scores with Claude Code compared to PI; for example, Opus 5.5 achieved a 4.8-point gain with Claude Code over PI. In contrast, using PI instead of Codex led to gains of 7.6, 1.9, 1.9 and 8.6 points for GPT-6 Astra, GPT-6.1 Sol, GPT-6.1 Terra and GPT-5.6 Sol respectively.

Increased resource usage did not translate to improved performance (Figure 2B–E). Gemini 3.8 Flash, DeepSeek V4.1 Flash and GLM-5.3 scored 2.9%, 0% and 0% despite averaging 37M, 60M and 38M tokens and 164, 269 and 244 tool calls per run, against 21M tokens and 92

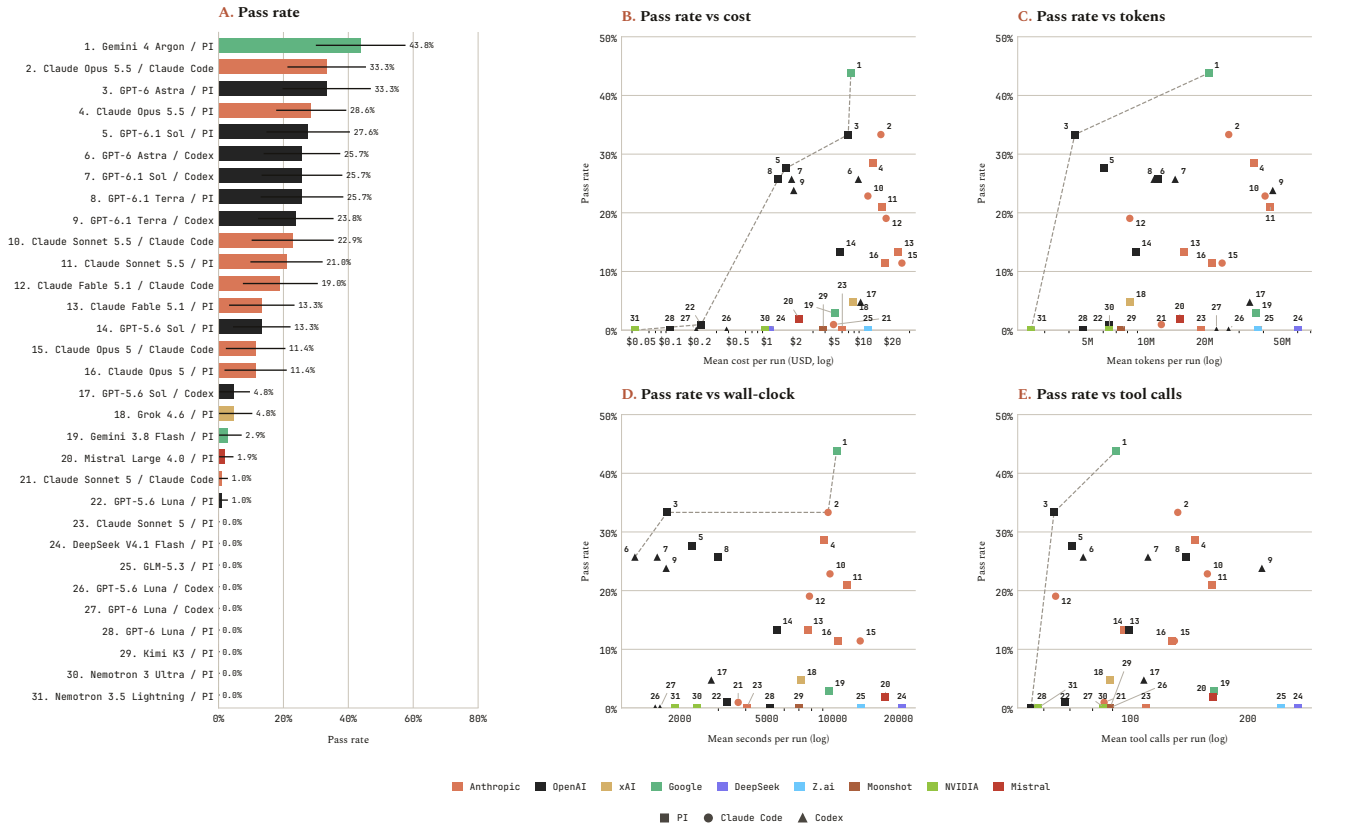


Figure 2: Pass rate by configuration. (A) Mean pass rate over the 35 tasks, three attempts each, with 95% Student's t intervals taken over per-task pass rates. Pass rates are computed over all three attempts per task, counting timeouts, memory crashes, and refusals as failures. All Claude runs were executed through the LSVP program. (B–E) Pass rate against mean cost, total tokens, wall-clock time and tool calls per run, on log scales. Cost, tokens and tool calls are means over the runs with available usage; wall-clock is the mean over all runs. Points are numbered by their rank in (A). The dashed line marks the Pareto frontier.

tool calls for Gemini 4 Argon.

In fact, the top-performing configurations also tended to be the most efficient. GPT-6 Astra / PI reached the same pass rate as Opus 5.5 / Claude Code at 2.2 times lower cost, 6.2 times fewer tokens, 5.4 times less wall-clock time and 2.1 times fewer tool calls per run. GPT-6 Astra / PI used $6.95, 4.3M tokens, 29 minutes and 64 tool calls per run against $5.73, 8.8M tokens, 92 minutes and 99 tool calls for GPT-5.6 Sol / PI (13.3%), and Opus 5.5 / Claude Code used 26.5M tokens, 159 minutes and 132 tool calls per run at $15.14 against 40.8M tokens, 162 minutes and 157 tool calls for Sonnet 5.5 / Claude Code (22.9%). Harness choice also changed efficiency independently of accuracy: GPT-6 Astra used 2.6 times fewer tokens on PI than on Codex while scoring higher.

Performance by Category

Mean pass rates across the fourteen non-zero scoring models were highest for tasks involving variants in snRNA genes, cases with single homozygous variants, and *de novo* autosomal dominant inheritance. They were lowest for somatic or non-genetic cases (cases without a germline cause) and autosomal dominant inherited causes (Figure 3). Pass rates in these categories sometimes grouped by model series: Opus 5.5, Sonnet 5.5, and Fable 5.1 all scored best on single homozygous variants and the three GPT-6 series models all scored best on compound heterozygous variants.

Gemini 4 Argon had the highest overall pass rate and led 16 of the 21 task categories (Figure 3). Claude models led the other five: single homozygous variants, dual diagnoses, synonymous or splice-region variants, cases affecting the cardiovascular system or exhibiting congenital anomalies, and inherited autosomal dominant cases, where Fable 5.1 tied with Grok 4.6. Only 7 models exhibited a non-zero pass rate for non-genetic cases.

Global Failure Modes

We assigned every failed run one primary failure mode and any number of contributing failure modes from six classes, and report shares over the 2097 failed runs of the 14 models that passed at least one attempt unless stated otherwise (Figure 4). *Execution and tool use* covers runs that timed out, mismanaged their compute, or failed to use tools correctly. *Data acquisition and work-up* covers runs that failed to obtain needed data, including sequencing data, stains, imaging studies, functional assays, or lab tests. We also include runs that closed the case prematurely or failed to recover from denials. *Search design* covers runs that held the right data but could not find the right variant in it. *Interpretation* covers runs that reached the variant but reasoned wrongly about it. *Answer completeness* covers runs that identified a variant but missed required evidence, a second component of a dual diagnosis, or were penalised for over-inclusion.

Data acquisition and work-up was the largest failure class among the models that passed at least one attempt (Figure 4A). It was the primary cause in 36% of failed runs, ahead of search design (29%) and interpretation (17%). Execution faults, mostly wall-clock timeouts, were the primary cause in 13% of failures. 70% of Mistral Large 4.0 runs, 29% of Gemini 4 Argon runs and 21% of Claude Opus 5 runs timed out, against none of the 210 runs each of GPT-6 Astra, GPT-6.1

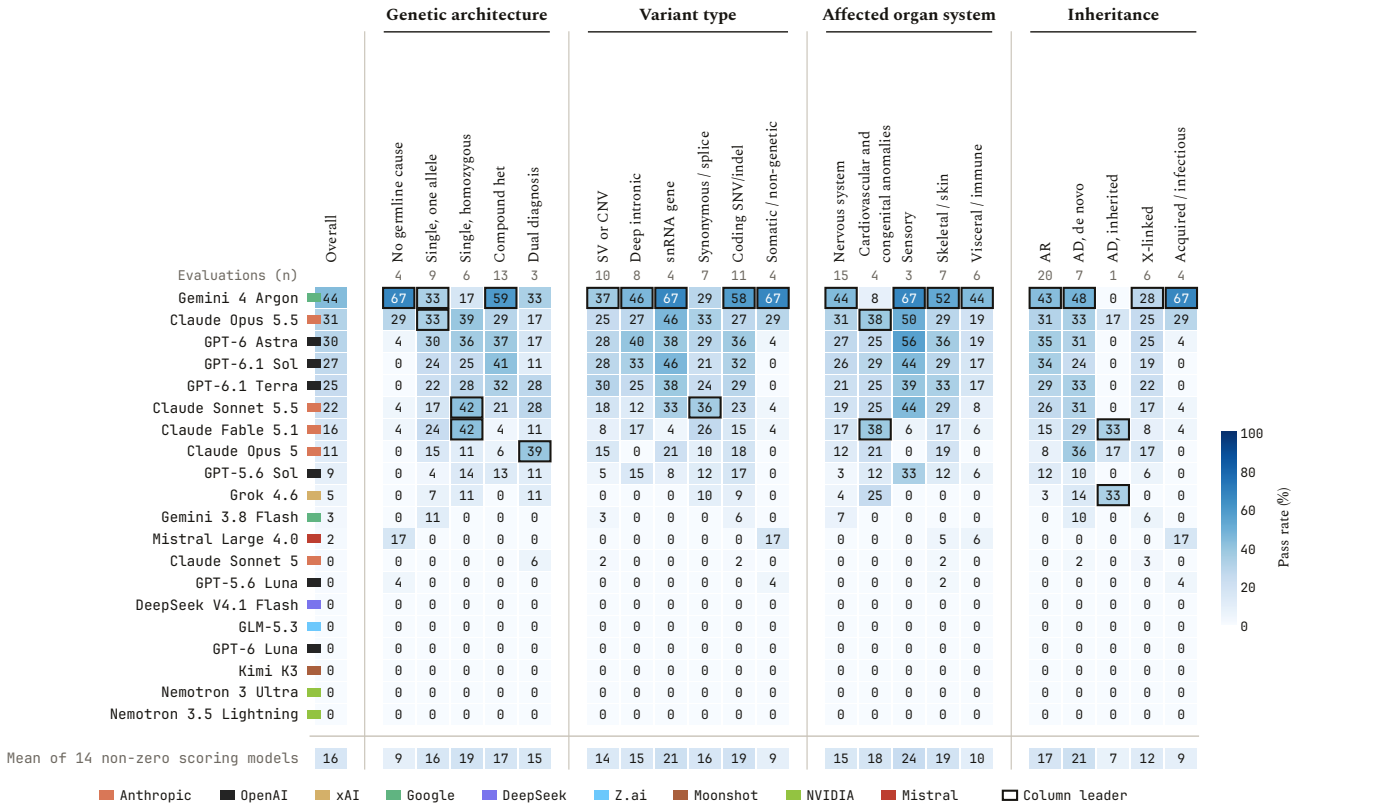


Figure 3: Pass rate by genetic architecture, variant type, affected organ system and inheritance. All 20 models, harnesses pooled, ordered by overall pass rate; runs that timed out or crashed count as failures, as in Figure 2. The outlined cell is the leader of each column and the bottom row is the mean over the models that passed at least one attempt. Abbreviations: het, heterozygous; SV, structural variant; CNV, copy-number variant; SNV, single-nucleotide variant; snRNA, small nuclear RNA; AR, autosomal recessive; AD, autosomal dominant.

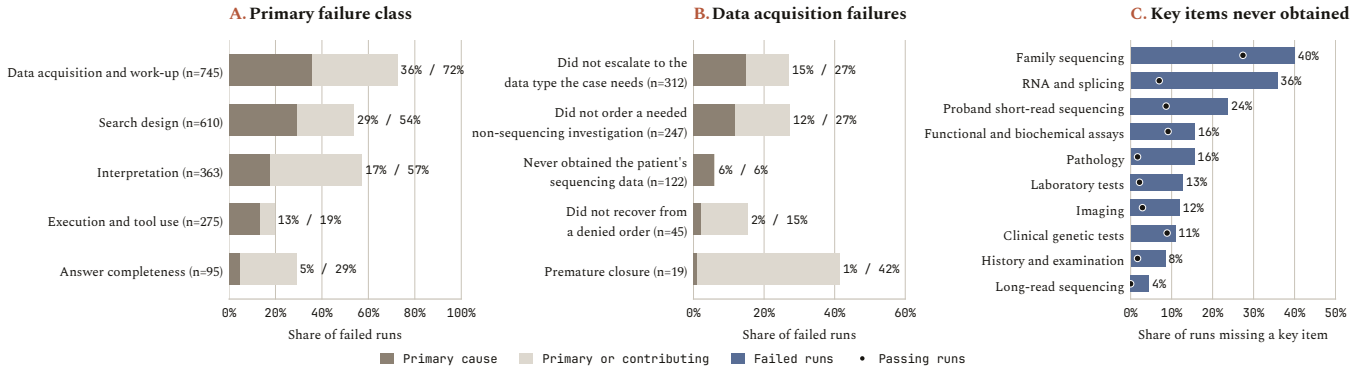


Figure 4: Failure modes of the models that passed at least one attempt. Each of 2097 failed runs in models with a non-zero pass rate was assigned one primary failure mode and any number of contributing failure modes. (A) Share of failed runs whose primary cause falls in each class (dark), and the share that carry the class as primary or contributing (light). (B) The same for subcategories of data acquisition and work-up. (C) Share of failed runs that never obtained at least one key data object of each type, against the same share among passing runs (dots).

Sol and GPT-6.1 Terra. Compute or resource mismanagement contributed to 15% of failures. Execution faults were the primary cause of 25% of failures among models with no passing runs, and contributed to 46%. DeepSeek V4.1 Flash timed out in 84% of its runs, largely due to inefficient data processing. Interpretation errors contributed to 57% of failures, most often as anchoring on a wrong diagnosis or gene (38%).

Within data acquisition, the most common primary failure was not escalating to the data type the case needed (15% of failed runs), followed by not ordering a needed non-sequencing investigation such as a stain, imaging study or laboratory test (12%); only 6% never obtained the patient's sequencing data at all (Figure 4B). Failed runs most often went without family sequencing (40%), RNA or splicing studies (36%) and the proband's short-read data (24%) (Figure 4C).

Case Study Failure Modes

To further illustrate agent behavior on RareDiseaseBench, we provide five case studies and analysis of frontier model behavior on each. We describe Gemini Argon 4 behavior in detail in the last case only because it timed out in all others.

Case study 1: a 13-year-old boy with variants in *SAMHD1* and *MN1*. In this case, a boy presents with destructive chilblain-like lesions

with autoamputation of toes, polyarthritis, growth restriction, microcephaly and sensorineural hearing loss [19]. There are two genetic causes: a *de novo MN1* nonsense variant that explains the syndromic features and is visible in exome sequencing data, and a homozygous balanced translocation through *SAMHD1* intron 4 that causes the interferonopathy and which is not detectable using alignment to GRCh38. To pass, an agent must see that *MN1* leaves the chilblains and interferon signature unexplained, request long reads, and perform a T2T-CHM13 alignment. Opus 5.5 submitted *MN1* as the only cause while noting that it "is not an established interferonopathy gene." GPT-6 Astra runs ordered karyotyping and long reads and reported the *SAMHD1* translocation, but sometimes missed *MN1* entirely.

Case study 2: a 21-year-old man with a deep intronic *ZIC3* variant. In this case, a man presents with heterotaxy with right atrial isomerism, which typically points to maternally-inherited variants in *ZIC3* [49]. The cause here is a variant 249bp past the 3′ end of the MANE transcript, annotated as downstream with no SpliceAI score, that an alternative transcript places in an intron where it creates a 57-bp pseudoexon capturing the exon 2 donor both transcripts share. To pass, an agent should examine rare non-coding variants instead of filtering on the MANE annotation, score splicing on the other transcript, and order functional splicing studies to investigate the mechanism. GPT-6 Astra consistently reported the variant in five of six runs but never requested the splicing studies. In contrast, Opus 5.5 passed twice by requesting functional studies on its splicing hypothesis, once with the full pseudoexon mechanism and once with minigene assay evidence.

Case study 3: a boy with a *de novo* duplication of *TAB3*. A boy presents with progressive thinning of scalp hair with preserved eyebrows and body hair, and a biopsy shows follicular miniaturization [24]. The cause is a *de novo* 357-kb tandem duplication at Xp21.2 containing *TAB3*. *TAB3* has no prior disease associations, but data is available for request involving scalp staining, a mouse model, and keratinocytes. To pass, an agent must find the duplication, argue for *TAB3* over its neighbours from its role in NF-κB signaling, and order the functional studies that test the link. Opus 5.5 could precisely describe the duplication but tended to dismiss it as a VUS "with no hair link." GPT-6 Astra and Argon runs instead claimed to a *de novo* missense in *MAB21L4* that is native to the synthetic genome. Both models were capable of passing, however, if they deeply pursued *TAB3*, showing themselves capable in one out of 96 runs to order scalp staining.

Case study 4: a 16-year-old boy with compound heterozygous *RNU2-2* variants. In this case, a 16-year-old boy of average growth presents with a language disorder, learning disability and autism [31]. The cause is a compound heterozygous pair in *RNU2-2*, a small nuclear RNA gene the exome does not cover and that is known for a dominant *de novo* disorder. As *RNU2-2* variants are easy to identify for models, we incorporated a *de novo DNMT3A* missense variant in the genome to act as a decoy (in the donor from whom the synthetic genome was made, this did not cause disease). To pass, an agent must discover the two inherited *RNU2-2* variants and reject *DNMT3A* on the absence of an overgrowth phenotype. Every completed Opus 5.5 and GPT-6 Astra run submitted *DNMT3A*: Opus diagnosed Tatton-Brown-Rahman syndrome "without overgrowth", while Astra called the diagnosis inconclusive. Two Opus runs found both *RNU2-2* variants and discarded them.

Case study 5: a boy with a complex *DMD* rearrangement. In our final case, a boy has delayed motor milestones, calf hypertrophy, Gowers' sign and a creatine kinase of 10,100 U/L. This case features two visits: an initial diagnosis followed by a judgment on an outside laboratory's report of a *DMD* inversion [50]. The patient's long-read sequencing data shows a complex rearrangement in which segments containing exons 1 and 2 are duplicated and joined head to head, contradicting, in fact, the inversion claim made in the case report this task was drawn from. To pass, an agent must order targeted long-read sequencing at the first visit, then show how the conclusion from the outside laboratory is wrong. Opus 5.5 resolved the full structure, while Astra and Argon runs ordered the long reads but refused to comprehensively refute the outside laboratory's claim based on read evidence. Argon further failed to elaborate that one reason the inversion might not be real is the presence of a large number of reference-spanning reads inconsistent with the inversion structure.

Discussion

When constructing RareDiseaseBench, we often found that agents are already capable of analyzing many rare disease cases, especially documented coding variants in whole exome or whole genome sequencing data. However, they were significantly less capable of analyzing cases involving multiple or complex variants and performed poorly when we asked them to reason about what assay might be useful next. RareDiseaseBench shows that agents powered by frontier models do not yet have the ability to reliably request, analyze, and interpret data to diagnose challenging rare disease cases. The highest overall pass rate was 43.8%, with other frontier models trailing at 33%. Agents display inefficient use of time, tool calls, tokens, and compute resources. The design of RareDiseaseBench allows us to observe that agents often fail to retrieve the right data, and even when they do, fail to integrate findings correctly with previous findings.

RareDiseaseBench has several limitations. First, RareDiseaseBench is based off of publicly available cases using synthetic genomes. This introduces risks of contamination and potentially reduces how well RareDiseaseBench generalizes to real patients. Second, RareDiseaseBench makes use of LLM-as-a-judge systems for rubric grading and Records release. While we observed 98.1% unanimous agreement across judges in 15,879 judgements across all tasks, observed little to no reward-hacking behavior, and enforced sidecar tests for all data objects, we cannot exclude the possibility that grading and data release decisions were susceptible to misaligned agent behavior or were disconcordant with human expertise. Finally, the design of RareDiseaseBench imposes restrictions on agent behavior: agents must request data one assay at a time, sometimes with fuzzy semantic justification; agents can accidentally submit answers; and agents are restricted in their compute and time resources. Especially for models at the frontier, we observe failure modes that we believe if addressed would be consequential for rare disease patients; even so, we acknowledge that in some capacity RareDiseaseBench tests agents' ability to learn and make use of our Records system. Future work could address these limitations by using unpublished patient data and improving the RareDiseaseBench harness to better proxy real-world settings, such that measuring model ability in the system is a feature, not a bug.

Thus, we hope that RareDiseaseBench will inspire future work on ever-more-realistic scientific and clinical benchmarks. In the present, however, we believe that RareDiseaseBench provides a useful barometer for model makers, patients, clinicians, and researchers. We fervently hope that advancements in AI as measured or motivated by this benchmark will hasten time to diagnosis for all rare disease patients.

Acknowledgements

We thank the clinicians who provided inspiration and feedback on this benchmark. We thank many members of the scientific and engineering staff at LatchBio for their support and ideas. We thank the patients and families whose case studies inspired our benchmark tasks.

References

1. The Lancet Global Health. The landscape for rare diseases in 2024. Lancet Glob Health 12(3):e341, 2024. doi:10.1016/S2214-109X(24)00056-1.
2. Valdez R, Ouyang L, Bolen J. Public health and rare diseases: oxymoron no more. Prev Chronic Dis 13:150491, 2016. doi:10.5888/pcd13.150491.
3. Gürkan H, Bilge Satkın N. The importance of genetic diagnosis in rare diseases. Balkan Med J 42(2):92–93, 2025. doi:10.4274/balkanmedj.galenos.2025.2025-270125.
4. Splinter K, Adams DR, Bacino CA, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. N Engl J Med 379(22):2131–2139, 2018. doi:10.1056/NEJMoa1714458.
5. The 100,000 Genomes Project Pilot Investigators. 100,000 Genomes pilot on rare-disease diagnosis in health care—preliminary report. N Engl J Med 385(20):1868–1880, 2021. doi:10.1056/NEJMoa2035790.
6. Zhao W, Wu C, Fan Y, Qiu P, Zhang X, Sun Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature 651(8106):775–784, 2026. doi:10.1038/s41586-025-10097-9.
7. Zou C, Banerjee A, Bhasin H, Muralidharan HS, Yang Z, Diks IF, Seeyave E, Wang D, Le HGBH, Proctor T, Workman K. OmicsBench: A large-scale, verifiable benchmark for agentic life sciences capabilities. latch.bio, 2026. `https://latch.bio/omicsbench`
8. Le H, Ramasamy R, Urrutia A, Yazdani M, Proctor T, Workman K. TxBench-PP: Analyzing AI agent performance on small-molecule preclinical pharmacology. arXiv [cs.AI], 2026.
9. Ramasamy R, Le H, Brougher J, Urrutia A, Banerjee A, Poust S, Workman K. TxBench–Antibody Discovery: Benchmarking AI agents on experimentally grounded decisions in therapeutic antibody discovery. benchmarks.bio, 2026.
10. Badshah J, Banerjee A, Liu Y, Eliasen A, Jain A, Proctor T, Flood D, Workman K. TxBench–Antibody Development: Benchmarking AI agents on the decisions that take a therapeutic antibody from lead nomination to the first human dose. latch.bio, 2026.
11. Badshah J, Saye BK, Banerjee A, Shrishrimal S, Ramasamy R, Flood D, Workman K. OncologyBench: Benchmarking multiagent AI systems on grounded systems in oncology research and treatment. latch.bio, 2026.
12. Soskin Hicks R, Trofimov M, Lim D, Arora RK, Tsimpourlas F, Bowman P, et al. HealthBench Professional: Evaluating large language models on real clinician chats. arXiv:2604.27470 [cs.CL], 2026.
13. Jaech A, Cheatham M, Shringarpure SS, Genetti CA, Pradhan P, Bagul A, et al. LLM-assisted reanalysis of unsolved rare disease genomes increases diagnostic yield. NEJM AI 3(7), 2026. doi:10.1056/AIcs2501343.
14. Ning J, Lin J, Fang Y, Li W, Liu J, Tang C, et al. MMRareBench: A rare-disease multimodal and multi-image medical benchmark. arXiv:2604.10755 [cs.CV], 2026.
15. McKinnon D. RareBench 0.1: Benchmarking agentic AI for rare disease diagnosis. Gamow Labs, 2026. `https://gamowlabs.com/rarebench-0-1`
16. Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments. Software, 2026. doi:10.5281/zenodo.20953922. `https://harborframework.com/`
17. Alabdi L, Cogne B, Almasood AS, Alsehly A, Helaby R, Maddirevula S, et al. WDR59 Is Mutated in Individuals With Autosomal Recessive Syndromic Dilated Cardiomyopathy. Clinical Genetics 110(1):84–89, 2026. doi:10.1111/cge.70151.
18. Arriaga TM, Mendez R, Ungar RA, Bonner DE, Matalon DR, Lemire G, et al. Transcriptome-wide outlier approach identifies individuals with minor spliceopathies. The American Journal of Human Genetics 112(10):2458–2475, 2025. doi:10.1016/j.ajhg.2025.08.018.
19. Baker PJ, Zhang Y, Bishop I, Cleveland ML, McAllan AL, Hollway GE, et al. Iterative genetic testing identifies SAMHD1 deficiency caused by a homozygous balanced translocation. Journal of Human Immunity 2(6):e20260044, 2026. doi:10.70962/jhi.20260044.
20. Besse A, Wu P, Bruni F, Donti T, Graham BH, Craigen WJ, et al. The GABA Transaminase, ABAT, Is Essential for Mitochondrial Nucleoside Metabolism. Cell Metabolism 21(3):417–427, 2015. doi:10.1016/j.cmet.2015.02.008.
21. Bilgrav Saether K, Eisfeldt J, Bengtsson JD, Lun MY, Grochowski CM, Mahmoud M, et al. Leveraging the T2T assembly to resolve rare and pathogenic inversions in reference genome gaps. Genome Research 34(11):1785–1797, 2024. doi:10.1101/gr.279346.124.
22. Bonde LD, Holling T, Alawi M, El Beheiry AA, Mir Hassani Z, Bachand F, et al. A homozygous synonymous NOP58 variant causes a neurodevelopmental disorder by impairing maturation of pre-ribosomal RNAs. Human Genetics and Genomics Advances 7(1):100557, 2026. doi:10.1016/j.xhgg.2025.100557.
23. Calvo M, Reynolds G, Luca M, Di Gregorio E, Cardaropoli S, Salvo E, et al. TRAPPC9-Related Intellectual Developmental Disorder: A Systematic Review and a Novel Case of a Complex Structural Variant. Genes 17(6):658, 2026. doi:10.3390/genes17060658.
24. Cao Q, Zhao A, Wang J, Lin X, Long Z, Pan C, et al. Overlapping Xp21.2 duplications define an X-linked hypotrichosis simplex and implicate TAB3 dosage sensitivity. The American Journal of Human Genetics 113(8):1719–1735, 2026. doi:10.1016/j.ajhg.2026.06.013.
25. Cavestro C, Panteghini C, Reale C, Nasca A, Fenu S, Salsano E, et al. Novel deep intronic mutation in PLA2G6 causing early-onset Parkinson's disease with brain iron accumulation through pseudo-exon activation. neurogenetics 22(4):347–351, 2021. doi:10.1007/s10048-021-00667-0.
26. Chen J, Ruxian A, Li D, Jiang Y, Ma B. Erdheim–Chester Disease with Breast and Axillary Involvement Diagnosed by Ultrasound-Guided Biopsy: A Case Report and Literature Review. Journal of Clinical Medicine 15(16):6330, 2026. doi:10.3390/jcm15166330.
27. Feng Y, Yang L, Xu QB, Cao LF. Central precocious puberty as the initial manifestation of multisystem involvement caused by de novo heterozygous KMT2B mutation and STS hemizygous deletion: a case report. Frontiers in Endocrinology 17:1869758, 2026. doi:10.3389/fendo.2026.1869758.
28. Fernández-Eulate G, Duval R, Konyukh M, Bertini A, Candayan A, Jordanova A, et al. Biallelic PIGB Variants Cause Motor Neuropathy with Conduction Blocks and Peripheral Nerve Hyperexcitability. Annals of Neurology ana.78358, 2026. doi:10.1002/ana.78358.
29. Fröhlich J, Testa F, Kaminska K, Karali M, Zeuli R, Cancellieri F, et al. CLUAP1 variants cause non-syndromic retinitis pigmentosa. Graefe's Archive for Clinical and Experimental Ophthalmology, 2026. doi:10.1007/s00417-026-07469-w.
30. Giba N, Dunkel KO, Boros E, Csomor J, Paróczai D, Lantos T, et al. Whipple Disease with Concurrent Mycobacterium szulgai Infection: A Rare Diagnostic and Therapeutic Challenge. Diagnostics 16(15):2315, 2026. doi:10.3390/diagnostics16152315.
31. Greene D, Mendez R, Lees J, Barbosa M, Bruselles A, Chiriatti L, et al. Biallelic variants in RNU2-2 cause the most prevalent known recessive neurodevelopmental disorder. Nature Genetics 58(4):774–781, 2026. doi:10.1038/s41588-026-02539-5.
32. Hegde P, Anand A, Rani R, Kaur N, Shah A, Kulkarni S, et al. Biallelic Loss-Of-Function Variant in ATP5ME Is Associated With Severe and Early Onset Oxidative Phosphorylation Deficiency. Journal of Inherited Metabolic Disease 49(4):e70222, 2026. doi:10.1002/jimd.70222.
33. Hirsch Y, Tangshewinsirikul C, Booth KT, Azaiez H, Yefet D, Quint A, et al. A synonymous variant in MYO15A enriched in the Ashkenazi Jewish population causes autosomal recessive hearing loss due to abnormal splicing. European Journal of Human Genetics 29(6):988–997, 2021. doi:10.1038/s41431-020-00790-w.
34. Hnízda A, Martinez-Delgado B, Sanchez-Ponce D, Alonso J, Amiel J, Attie-Bitach T, et al. De novo EHMT2 variants cause an autosomal dominant EHMT2-related Kleefstra syndrome via loss of G9a methyltransferase activity. Nature Communications 17(1):8203, 2026. doi:10.1038/s41467-026-74987-w.
35. Horowitz-Cederboim S, Hoffman-Lipschuetz R, Durst R, Shpitzen S, Shauer A, Zwas DR, et al. Reconsidering a silent variant: SGCA's role in atypical cardiomyopathy. European Journal of Human Genetics 34(4):528–534, 2026. doi:10.1038/s41431-025-01981-z.
36. Hu T, Wang C, Liu C, Sun J. Hepatobiliary langerhans cell histiocytosis mimicking Caroli syndrome and recurring in a liver allograft. Frontiers in Medicine 13:1898715, 2026. doi:10.3389/fmed.2026.1898715.
37. Kennedy AD, Pappan KL, Donti T, Delgado MR, Shinawi M, Pearson TS, et al. 2-Pyrrolidinone and Succinimide as Clinical Screening Biomarkers for GABA-Transaminase Deficiency: Anti-seizure Medications Impact Accurate Diagnosis. Frontiers in Neuroscience 13:394, 2019. doi:10.3389/fnins.2019.00394.
38. Lin S, Cancellieri F, Cao Y, Lotery AJ, Moye AR, Vaclavik V, et al. Bi-allelic variants in FSD1L cause retinitis pigmentosa with or without neurological involvement. The American Journal of Human Genetics 113(3):616–626, 2026. doi:10.1016/j.ajhg.2026.01.015.
39. Pitsava G, Bluske K, Barrick R, Dios ID, Duong C, Blanco K, et al. Yield of Long-Read Genome Sequencing for Rare Disease Diagnosis in Short-Read Genome Negative Cases. medRxiv [preprint], 2026. doi:10.64898/2026.09.09.26362331.
40. Rabiei S, Collins D, Schwartz P, Pallante A, Song X, Kantorovich V. Intracranial phosphaturic mesenchymal tumor mimicking meningioma and causing tumor-induced osteomalacia: a case report. JBMR Plus 10(11):ziag156, 2026. doi:10.1093/jbmrpl/ziag156.
41. Rius R, Blakes AJM, Chen Y, De Jonghe J, Lecoquierre F, Dawes R, et al. Biallelic variants in the noncoding RNA gene RNU4-2 cause a recessive neurodevelopmental syndrome with distinct white matter changes. Nature Genetics 58(4):761–773, 2026. doi:10.1038/s41588-026-02554-6.
42. Rossi V, Brooks D, Dai H, Mizerik E, Salazar K, Davila-Williams D, et al. A Rare Molecular Diagnosis in a Patient With Hepatocerebral Syndrome Contributes to the Expansion of the Phenotypic Spectrum of POLG2 -Related Mitochondrial Disorder. American Journal of Medical Genetics Part A 197(11):e64177, 2025. doi:10.1002/ajmg.a.64177.
43. Scheidecker S, Bär S, Kröll-Hermi A, Delvallée C, Rinaldi B, Korpioja A, et al. Novel MYH10 heterozygous variants associated to a syndrome combining mainly ptosis and ocular coloboma expand the MYH10 related phenotypes. European Journal of Human Genetics 33(11):1432–1441, 2025. doi:10.1038/s41431-025-01803-2.
44. Simon MT, Eftekharian SS, Ferdinandusse S, Tang S, Naseri T, Reupena MS, et al. ECHS1 disease in two unrelated families of Samoan descent: Common variant - rare disorder. American Journal of Medical Genetics Part A 185(1):157–167, 2021. doi:10.1002/ajmg.a.61936.
45. Sneddon TP, Melville SA, Xiong M, Foss K, Cardwell EB, Rafferty K, et al. Long-Read Sequencing of a Neurodevelopmental Disorder Patient Reveals Complex Rearrangement Involving the ARID1B Gene. American Journal of Medical Genetics Part A 200(2):426–431, 2026. doi:10.1002/ajmg.a.64249.
46. Spranger S, Faust H, Duffek P, Schubert S, Skladny H, Kovacs P, et al. Case Report: Deep intronic PHEX variant causing aberrant splicing identified by whole

genome and targeted RNA sequencing in X-linked hypophosphatemia. Frontiers in Endocrinology 17:1847219, 2026. doi:10.3389/fendo.2026.1847219.

47. Sy PM, Baba M, Fujiwara K, Yamamura N, Kawato K, Nishi E, et al. A deep intronic IFT172 variant causing pseudoexon inclusion identified by whole-genome sequencing in nephronophthisis. CEN Case Reports 15(3):95, 2026. doi:10.1007/s13730-026-01132-3.

48. Tibbe D, Holling T, Spohn M, Alawi M, Nampoothiri S, Kutsche K. A homozygous synonymous variant in SMG5, encoding a nonsense-mediated mRNA decay factor, causes developmental delay with growth retardation and relative macrocephaly. Human Genetics 145(1):26, 2026. doi:10.1007/s00439-025-02813-2.

49. Wells JR, Padua MB, Haaning AM, Smith AM, Morris SA, Tariq M, et al. Non-coding cause of congenital heart defects: Abnormal RNA splicing with multiple isoforms as a mechanism for heterotaxy. Human Genetics and Genomics Advances 5(4):100353, 2024. doi:10.1016/j.xhgg.2024.100353.

50. Xie Y, Bao L, Yu X, Liu Y. Long-read sequencing identifies complex structural variants in DMD patients. BMC Medical Genomics 19(1):74, 2026. doi:10.1186/s12920-026-02352-3.

51. Yuen M, Zhang K, Marchant RG, Ishimura R, Graham M, Aung-Htut M, et al. A severe neurodevelopmental syndrome linked to a South Asian founder variant in the UFMylation adaptor CDK5RAP3. Acta Neuropathologica 151(1):48, 2026. doi:10.1007/s00401-026-03017-2.