NeuroscienceBench

Benchmarking AI agents on the ability to analyze and interpret neuroscience research data

Jackson Brougher[1], Anshika Agrawal[1], Tim Proctor[1], Martin Jacko[1]

[1]LatchBio, San Francisco, CA, USA

Correspondence: martin.jacko@latch.bio

Abstract

Neuroscience research increasingly depends on large datasets and novel experimental methods that demand specialized expertise. Here we introduce NeuroscienceBench, a verifiable benchmark that tests whether an AI agent can reach correct scientific conclusions from diverse neuroscience data. The benchmark comprises 91 evaluations, each focused on a research task testing a distinct skill. Agents receive the same data a wet- or dry-lab neuroscientist would and return a structured answer that is graded deterministically. Evaluations are classified along three axes (functional system, process class, biological level) and draw on 43 peer-reviewed studies spanning next-generation sequencing, electrophysiology, optical physiology, electroencephalography, brain-computer interface decoding, cell and tissue imaging, behavioral scoring, and clinical biomarkers. Across 34 model-harness configurations (21 models, 4 harnesses, 9,212 graded attempts) the mean pass rate was 21.0%. The strongest configuration, Claude Opus 5.5 with the Pi harness, passed 32.2% (95% CI 23.5–41.0). The type of data analyzed had the largest effect on performance, with mean pass rates ranging from 14% on less common omics readouts and 18% on clinical and biofluid data to about 25% on behavior and on EEG and sleep scoring. Accuracy rose with cost, but inexpensive configurations came within a few points. Current AI models can therefore help with some neuroscience analyses but cannot be relied on without expert judgment.

Introduction

Nervous system disorders are the leading cause of disability worldwide and account for a growing share of healthcare costs in high-income countries. As of 2025, they affected more than 40% of the global population [1]. Yet new treatments remain scarce. Only 5.9% of neurological and 7.3% of psychiatric drug programs entering Phase 1 reach approval, below the 7.9% average across all indications [2]. Many of these setbacks reflect gaps in our understanding of neurobiology. On the forefront are the researchers that face the challenge of large amount of data from cell and animal models that poorly recapitulate human disease, drug mechanisms that cannot produce a sufficient therapeutic effect, and clinical endpoints misaligned with disease heterogeneity and patient diversity.

Data acquisition is arguably no longer the main constraint on closing these gaps. Brain-wide cell atlases, regional and whole-brain connectomes, biobanks, genomic databases, population imaging studies, and electrophysiology archives have greatly expanded the data available to neuroscientists. The bottleneck has shifted from generating data to analyzing and interpreting it against an immense body of prior knowledge that no single scientist or research team can fully internalize. AI agents working alongside researchers could accelerate this work, but their reliability must be monitored to avoid propagating errors.

Existing benchmarks do not fully test these analytical and reasoning abilities. BrainBench showed that large language models (LLMs) outperform human experts at identifying which of two versions of an abstract reports a study's true results [3]. Such benchmarks measure how well a model has absorbed the existing literature, not whether it can analyze new data. Agentic biology benchmarks such as LAB-Bench [4] and BixBench [5], and our own SpatialBench [6], scBench [7], and EpiBench [8], do test analysis of experimental data, but none is built around the data types specific to neuroscience. Our disease-area benchmarks, TxBench [9][10][11][12], OncologyBench [13], and RareDiseaseBench [14], have shown that frontier agents can execute individual analyses competently and the failure is often one of reasoning rather than the actual execution. NeuroscienceBench applies this same design to neuroscience, where diverse data types and scientific decision making requires deep domain expertise.

Benchmark Construction

NeuroscienceBench comprises 91 evaluations drawn from 43 peer-reviewed publications and their public datasets, which we refer to as problems (Figure 1). A problem yields several evaluations when its dataset supports more than one independent decision. The evaluations cover the field broadly along all three taxonomy axes.

The functional-system axis records which nervous-system function is studied, such as motor, vision, olfaction & taste, somatosensation, language & social and others. The 33 cross-system evaluations address a tissue compartment or whole-person readout rather than a single pathway, most often cortical and forebrain neurons, vasculature, barriers and CSF, and glia and neuroimmune. The process-class axis separates basic-science questions about how the nervous system normally works from questions about how it fails. Normal function is subdivided into eight subcategories, such as brain state and sleep, cognition, and motor control, and disease, injury and aging into eleven process classes, such as neurodegeneration, infection, and aging. The biological-level axis records the scale at which the evaluation's decision is made, from molecular through cellular and circuit to tissue, organ, and organism behavior.

Each evaluation also carries a workflow-stage tag for one of six phases of the research cycle: hypothesis and target, model-system selection, experimental design, data processing, analysis and inference, and validation and translation. The underlying data include next-generation sequencing, extracellular electrophysiology, histology and cellular imaging, sleep-state scoring, BCI decoding, EEG and ECoG, behavioral scoring and pose tracking, patch clamp and multi-electrode arrays, and clinical and biofluid cohorts.

Evaluations were built with a multi-stage authoring pipeline. Before inclusion, each evaluation passed automated checks for structure, shortcuts (tasks answerable without analyzing the data), guessability and near-duplication. Grading is deterministic. Numeric answers must fall within a tolerance or admissible range, categorical answers must match exactly, and sets are scored by overlap.

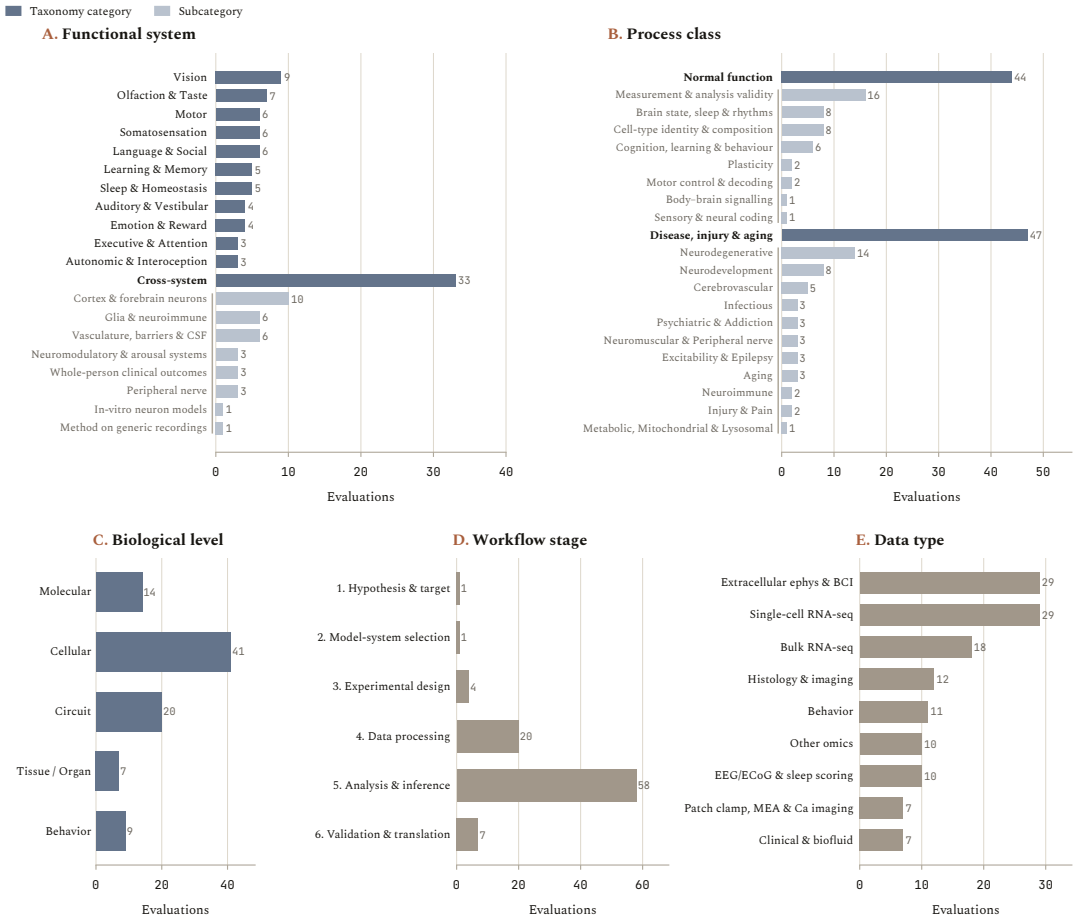


Figure 1: Benchmark anatomy. Evaluations per (A) functional system, (B) process class, (C) Biological level, (D) workflow stage, and (E) data type. Light bars in A and B are sub-categories: Cross-system is split by tissue compartment or readout, and process class is split into normal function and disease, injury and aging, then into sub-categories and disease classes. Each evaluation has one label on each taxonomy axis; an evaluation can use more than one data type, so E sums to more than 91.

Results

Leaderboard and efficiency

We evaluated 34 configurations of 21 models under four harnesses (Pi, Claude Code, OpenAI Codex and Grok Build), each making 3 attempts per evaluation without internet access, for 9,282 runs. After excluding platform failures (66) and refusals (4), 9,212 graded attempts remained.

No configuration passed more than 33% of attempts (Figure 2A). Claude Opus 5.5 with the Pi harness ranked first at 32.2% (95% CI 23.5–41.0), followed by Opus 5 / Pi (30.8%) and Sonnet 5.5 / Claude Code (29.7%). The top five configurations were all Claude models, and the confidence intervals of the top ten overlapped. Grok 4.7 / Grok Build (27.5%, sixth) was the strongest non-Anthropic configuration, and Nemotron 3.5 Lightning / Pi was the weakest (7.7%). The mean pass rate across all 34 configurations was 21.0%. The low ceiling does not reflect unsolvable tasks: 78 of the 91 evaluations were solved by at least one configuration, and 61 were solved on every attempt by at least one. The 13 evaluations that no configuration solved come from 10 problems, and include both pseudoreplication evaluations built on a glaucoma retina dataset and recording-level spike-sorter accuracy from SpikeForest.

Across the 29 configurations with reported cost, accuracy rose with spend (Spearman $\rho = 0.64$), but the relationship was loose (Figure 2B–E). GPT-6 Luna / Pi passed 20.4% of attempts at about $0.03 per run, within about 12 points of the leader at roughly one-hundredth of its cost, while the most expensive configuration, Fable 5.1 / Claude Code ($5.22 per run), passed 23.8%. Cost was not reported for five of the six Codex configurations, which are omitted from this comparison.

Across all configuration–evaluation pairs, 68.1% were never solved in three attempts, 11.4% were solved every time, and the remaining 20.5% were solved on some attempts but not others (Figure 2F). Even the leading configuration solved 42.9% of evaluations at least once, but only 23.1% on all three attempts.

The harness mattered for some models but did not favor any single design (Figure 2G). Opus 5 gained 5.1 points under Pi relative to Claude Code and Fable 5.1 gained 4.6 points, whereas Sonnet 5.5 did better under Claude Code (by 2.6 points), and Grok 4.7 did better under Grok Build than under Pi (by 4.0 points). For the six OpenAI models, Codex and Pi were within 2.6 points of each other in every case.

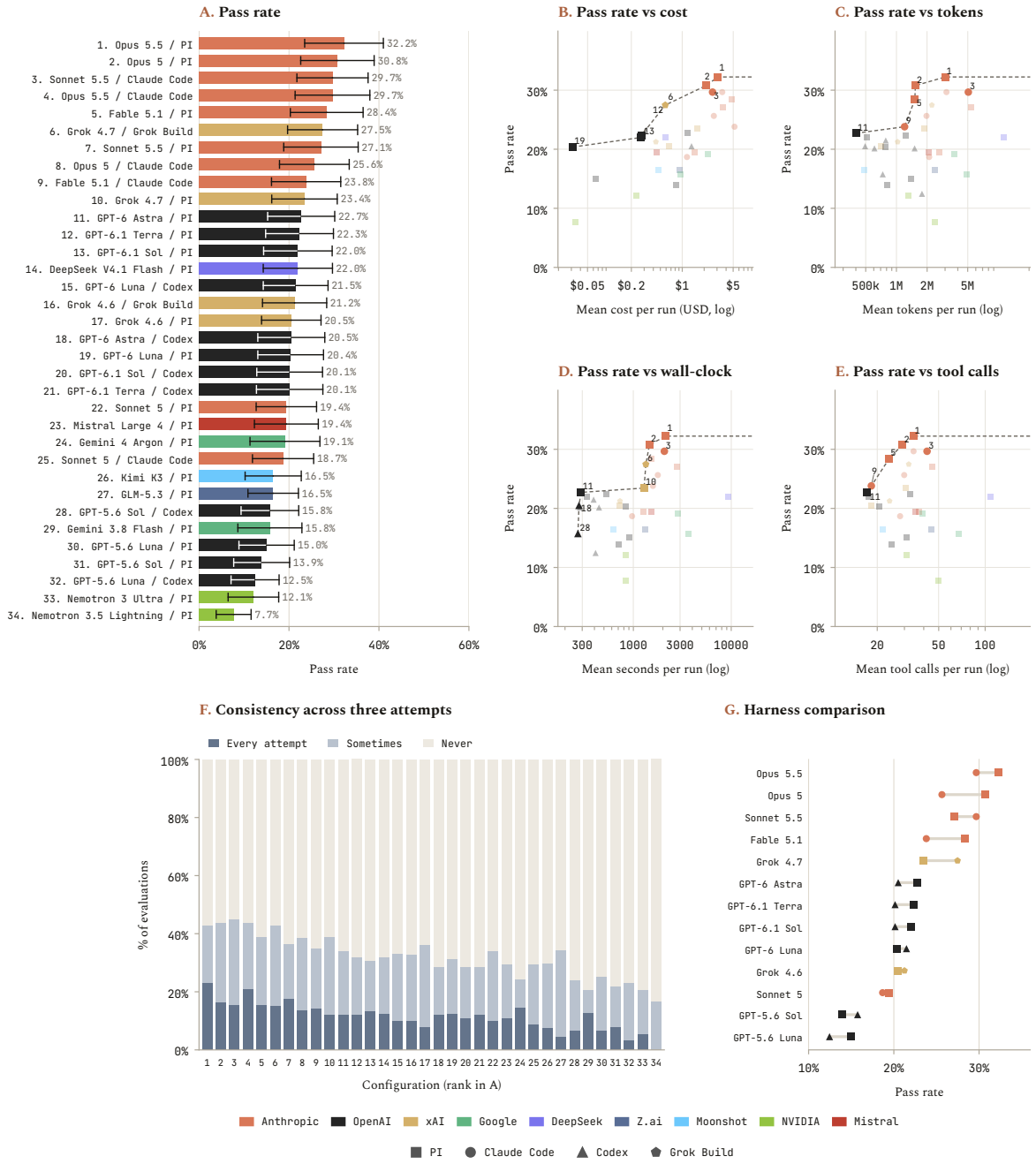


Figure 2: Leaderboard and run characteristics of 34 model–harness configurations. (A) Pass rate of each configuration over 91 evaluations and three attempts each, ordered by rank; whiskers are 95% Student's t intervals over per-evaluation pass rates and colors mark the model provider. (B–E) Pass rate against the mean cost (B), tokens (C), wall-clock time (D) and number of tool calls (E) per run, on log axes; marker shape gives the harness, numbers give the rank in A for the Pareto front and the top three configurations, and the dashed line traces the front. Cost was not reported for five of the six Codex configurations and tool calls for none of them, so those points are omitted from B and E. (F) For each configuration, the share of evaluations passed in every attempt, in some attempts, or never. (G) Pass rate of the 13 models run under two harnesses; the line joins the two configurations of a model. Refusals (4 runs) and platform failures (66 runs) are excluded throughout, leaving 9,212 graded attempts.

Performance by data type

Data type separated model performance sharply (Figure 3). Mean pass rates were lowest on the less common omics readouts, such as TCR-seq, metabolomics and proteomics (14.1%, 10 evaluations), and on clinical and biofluid data (17.6%, 7). The best performance was observed for EEG, ECoG and sleep scoring (25.9%, 10), histology and imaging (25.6%, 12), behavior and pose tracking (24.8%, 11) and bulk RNA-seq (24.6%, 18). The two most common data types, single-cell RNA-seq (16.3%, 29) and extracellular electrophysiology or BCI decoding (19.6%, 29), sat below the mean over models of 20.1%. Opus 5.5 passed 50.0% of bulk RNA-seq evaluations but none of those using patch clamp, multi-electrode arrays or calcium imaging, whereas GPT-6 Astra passed 35.7% of the latter and only 11.9% of clinical and biofluid evaluations. Across the 21 models, the gap between a model's best and worst pass rate averaged 28 points, indicating that the biological context carried by the data type strongly shapes agent performance.

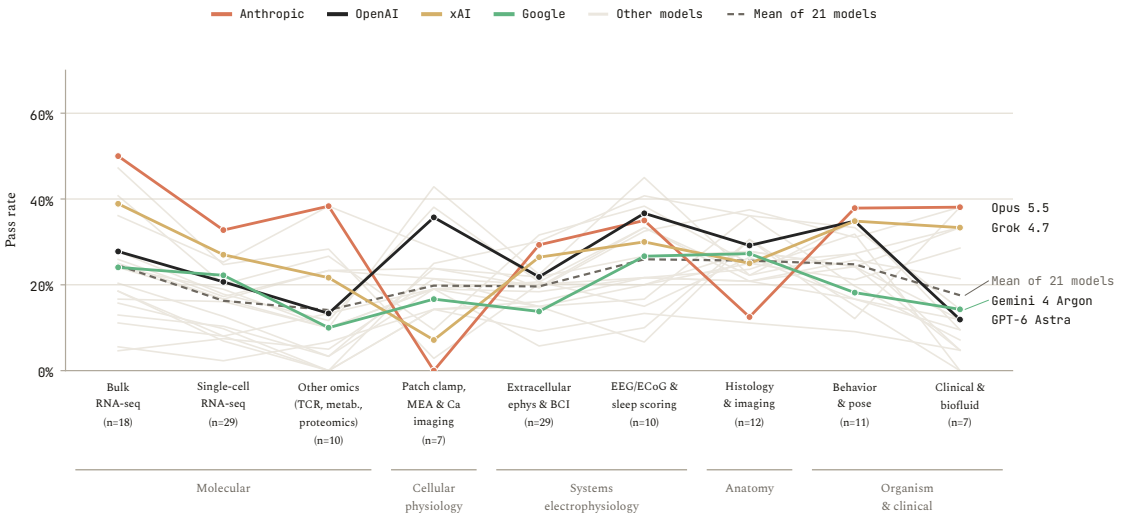


Figure 3: Pass rate by data type. Mean pass rate per data type, grouped by the biological scale of the data; the dashed line is the mean over the 21 models. An evaluation can use more than one data type; n is the number of evaluations using each. One flagship model from each lab in color and every other model in grey; both harnesses pooled and averaged at the evaluation level.

Performance by taxonomy

Agent accuracy varied widely across functional systems, from olfaction and taste (11.2%, mean over models), vision (11.3%) and auditory and vestibular (13.5%) to somatosensation (33.2%). Within cross-system, evaluations on cortical and forebrain neurons (13.2%, 10 evaluations) and on vasculature, barriers and CSF (14.3%) were harder than those on glia and neuroimmune cells (16.7%). This range should be read with caution, however. Most functional systems contain only three to nine evaluations, and pass rates vary as much within a system (pooled SD 17.9 percentage points across evaluations) as across the benchmark as a whole (17.8), so the differences between systems were not significant at the evaluation level (permutation test, $p = 0.50$) and are more likely to reflect which few evaluations happen to fall in each system than a property of the system itself.

By process class, agents were slightly more accurate on normal-function evaluations than on evaluations of disease, injury and aging (21.2% vs. 19.0%; Figure 4). Of the 21 models, 14 scored higher on normal function, as did 22 of the 34 configurations, but the gap was relatively small and not significant at the evaluation level (permutation test, $p = 0.53$).

By biological level, tissue- and organ-level evaluations were the hardest (13.0%, 7 evaluations), molecular (19.7%), cellular (19.7%) and circuit (21.2%) evaluations were intermediate, and behavioral evaluations were the easiest (25.1%).

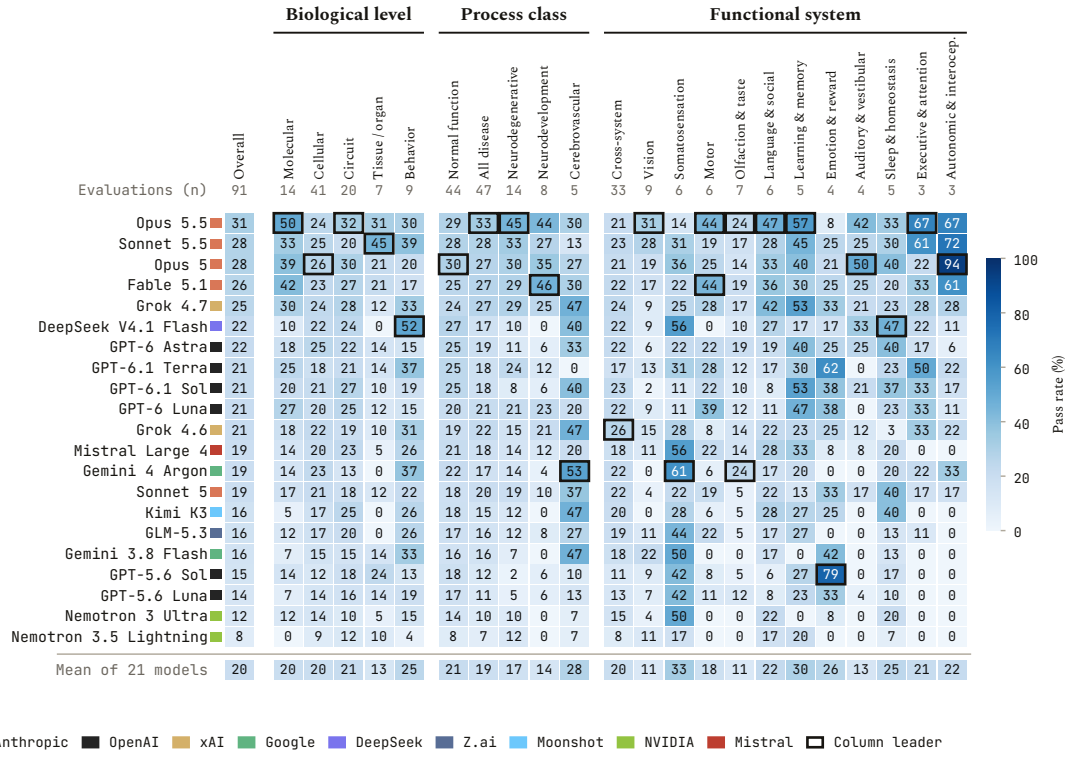


Figure 4: Pass rate by biological level, process class, and functional system. All 21 models, both harnesses pooled, ordered by overall pass rate. Each evaluation carries at most one biological level, process class, and functional system classification. The outlined cell is the leader of each column and the bottom row is the mean over models.

Failure decomposition by trap

Each evaluation was built around an analytical trap and has at least one inference fallacy the failing runs committed. The two describe the same design principle from different perspectives. A trap is indexed by data modality and names the processing or analysis choice on which an agent can go wrong. A fallacy is indexed by the form of the inference and names the faulty step from evidence to conclusion that the wrong choice produces. During authoring, every candidate evaluation was checked against both catalogs. A single evaluation therefore carries on average 2.6 trap categories and one or two fallacies. The 91 evaluations together covered 21 trap categories and captured 27 distinct fallacies.

Among traps (Figure 5A), quality control, controls and artifacts were the hardest (10.6%, 15 evaluations), followed by normalization and representation (12.6%, 8) and threshold or resolution choices (16.5%, 17). The three most common traps fell near the mean over models of 20.1%: analyzing the wrong subset or compartment (20.6%, 33), choosing the wrong replicate or unit of analysis (21.6%, 39) and computing a statistic that does not match the claim (22.2%, 32).

The hardest fallacies to avoid were a proxy read as the construct it stands for (4.5%, 3 evaluations), denominator neglect (7.9%, 3), equivocation (13.0%, 3), the continuum fallacy (13.4%, 5), scale or level confusion (13.7%, 7) and confounding or an omitted variable (14.8%, 20). Confounding is also the most common fallacy in the benchmark, present in more than one evaluation in five. Correlation read as causation (16.6%) and argument from ignorance (17.6%) followed, whereas the law of small numbers (20.0%), Simpson's paradox (24.5%), the significance fallacy (25.8%) and reification (26.0%) were intermediate, and the easiest fallacies were those of comparison and aggregation: a faulty baseline (33.9%), an inconsistent comparison (36.5%, 3 evaluations) and the ecological fallacy (37.1%, 4). Grouping the rows into statistical and reasoning classes gave similar overall rates (22.4% vs. 19.4%), so the difficulty lies in specific inference forms rather than in one class.

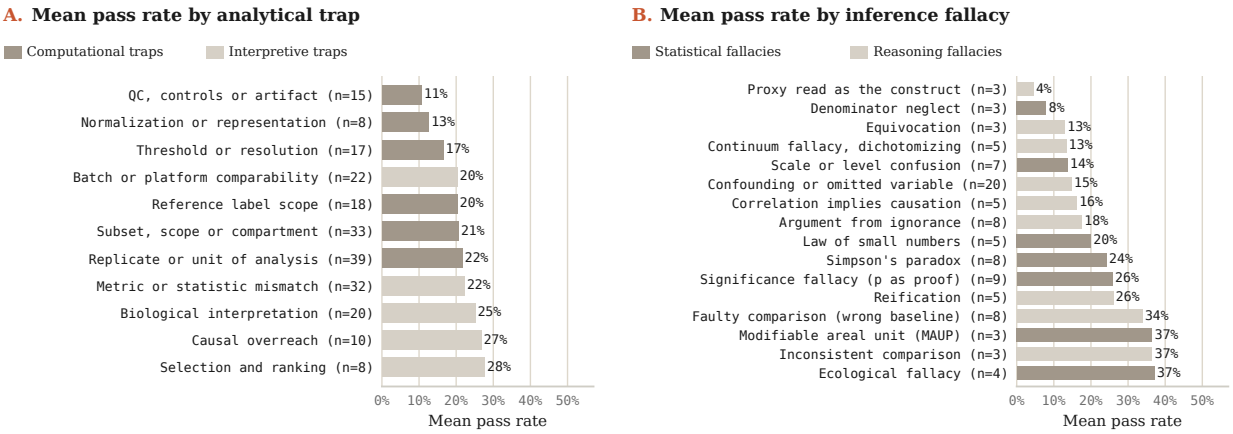


Figure 5: Mean pass rate by analytical trap and inference fallacy. (A) Analytical traps. (B) Inference fallacies. Bars give the mean over the 21 models of each model's pass rate on the evaluations in the row, both harnesses pooled, ordered from hardest to easiest; n is the number of evaluations.

Discussion

NeuroscienceBench provides a quantitative measure of how reliably AI agents perform the analytical and interpretive decisions that are common in neuroscience research. Currently, the strongest model configurations are Opus 5.5 and Opus 5 under Pi that passed 32.2% and 30.8% of attempts. Harness contributed less than model identity, yet the same model shifted by several points between harnesses. Although aggregate pass rates were similar across most categories, model rankings were not stable. The strongest category-dependent effect was data type: mean pass rates ranged from 14% on less common omics readouts and 18% on clinical and biofluid data to about 25% on behavior and on EEG and sleep scoring, and individual models swung by about 28% between their best and worst data type.

These results have several limitations. The evaluations are retrospective and built from published data; anti-shortcut checks reduce, but cannot exclude, the possibility that a model has encountered the source study. With 91 evaluations distributed over twelve functional systems and twelve process classes, many categories contain fewer than ten evaluations, and each configuration made only three attempts, so category-level estimates can be imprecise and should be treated as hypotheses for targeted expansion. Finally, initial difficulty was calibrated on two frontier models, Claude Opus 5.5 and GPT-6 Astra. The two models nevertheless fared very differently. Opus 5.5 ranked first overall, whereas GPT-6 Astra ranked eleventh under Pi and eighteenth under Codex.

References

1. World Health Organization. Global status report on neurology. Geneva: World Health Organization, 2025. ISBN 978-92-4-011613-9. link
2. Thomas D, Chancellor D, Micklus A, LaFever S, Hay M, Chaudhuri S, Bowden R, Lo AW. Clinical Development Success Rates and Contributing Factors 2011–2020. BIO, Informa Pharma Intelligence and QLS Advisors, 2021. link
3. Luo X, Rechardt A, Sun G, Nejad KK, et al. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour* 9(2):305–315, 2025. link
4. Laurent JM, Janizek JD, Ruzo M, Hinks MM, Hammerling MJ, Narayanan S, Ponnapati M, White AD, Rodriques SG. LAB-Bench: Measuring capabilities of language models for biology research. arXiv:2407.10362, 2024. link
5. Mitchener L, Laurent JM, Tenmann B, Narayanan S, Wellawatte GP, White A, Sani L, Rodriques SG. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. arXiv:2503.00096, 2025. link
6. Workman K, Yang Z, Muralidharan H, Le H. SpatialBench: Can agents analyze real-world spatial biology data? arXiv:2512.21907, 2025. link
7. Workman K, Yang Z, Muralidharan H, Abdulali A, Le H. scBench: Evaluating AI agents on single-cell RNA-seq analysis. arXiv:2602.09063, 2026. link
8. Muralidharan H, Baskar R, Lee SH, Proctor T, Workman K. EpiBench: Verifiable evaluation of AI agents on epigenomics analysis. arXiv:2606.13602, 2026. link
9. Le H, Ramasamy R, Urrutia A, Yazdani M, Proctor T, Workman K. TxBench-PP: Analyzing AI agent performance on small-molecule preclinical pharmacology. arXiv:2606.19245, 2026. link
10. 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. latch.bio, 2026. link
11. Jacko M, Brougher J, Urrutia A, Le H, Banerjee A, Flood D, Workman K. TxBench–Oligonucleotide Discovery. latch.bio, 2026. link
12. 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. link
13. 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. link
14. Chen Y, Banerjee A, Liu Y, Joshi P, Munagala N, Jacko M, Zou C. RareDiseaseBench: Can agents request, analyze, and interpret data to diagnose rare disease patients? latch.bio, 2026. link