NICHELEDGER: REPLACING REFERENCES AND GRADERS ON FIXED ANSWERS OF SPATIAL TRANSCRIPTOMICS AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Some scientific-agent benchmarks grade open-ended answers with a language model against a fixed reference. Replacing either can move a score on unchanged answers. Using NicheLedger, a spatial-transcriptomics agent benchmark, we replace each on fixed answers and track levels, orderings and test outcomes. On 36 niche-naming items under one grader, a frontier model's answers score 27.8% against clustering-derived references and 77.8% against clinician-consensus references, two physicians' hand grading gives 30.6% and 86.1%, and item-derangement controls score at most 8.3%. Of 18 gained items, seven clustering-derived references name a cell type no clinician named, three name no niche, and eight are wording, granularity or unresolved cases. One of 21 system-pair orderings reverses. The same model's 40 niche-naming answers score 30.0%, 47.5% and 20.0% under three grader states, two reached by names reporting one model, counting unparsable replies incorrect. Three claims each change outcome under some state. On BixBench, the ordering survives three graders and model-proposed references that two physicians prefer, which move accuracy by up to 4.5 and 1.4 points respectively. References and graders thus move levels and test outcomes while most orderings hold, and clinician consensus is not biological ground truth.

1 INTRODUCTION

Language-model agents increasingly perform end-to-end scientific analysis (Boiko et al., 2023; Wang et al., 2025b; Roohani et al., 2025), and their benchmarks (Mitchener et al., 2025; Laurent et al., 2024; Majumder et al., 2025; Workman et al., 2025; Liu et al., 2026) report aggregate success rates. For open-ended tasks, a language model typically compares each response with a fixed reference (Mitchener et al., 2025; Zheng et al., 2023). A score combines three inputs (candidate answers, a reference and a grader), only the first of which comes from the evaluated system. Interpreting it as capability requires checking the other inputs.

However, treating references and graders as fixed hides score changes on unchanged answers. Label errors can reorder top models when they are frequent enough (Northcutt et al., 2021), and corrected annotations reshape text-to-SQL leaderboards (Jin et al., 2026). A newly collected ImageNet test set costs models 11 to 14 points of accuracy (Recht et al., 2019), and re-annotated labels shrink reported gains (Beyer et al., 2020). Error-prone spatial-clustering annotations motivate expert-guided consensus (Sun et al., 2026). Hosted models drift behind unchanged names (Chen et al., 2024a; Leshin et al., 2026; Chauvin et al., 2026), and judge versions differ in strictness (Sunkavalli, 2026). Human-labeled anchors can separate judge drift from system drift (Li, 2026), but language-model judges disagree with domain experts more often than experts disagree with each other (Szymanski et al., 2025). The remaining question is which comparative conclusions survive a replacement.

To separate the answers from the two inputs that grade them, we introduce NicheLedger, a spatial-transcriptomics agent benchmark built from public histological slides (Jaume et al., 2024) with references from unsupervised clustering and a curated marker vocabulary. Replacing one input at a time on fixed answers attributes score changes to the replacement, not system improvement (Figure 1). Replacing the clustering-derived references for niche naming with the consensus of blinded clinicians, under one grader, raises the same frontier answers from 27.8% to 77.8%, while scoring each answer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

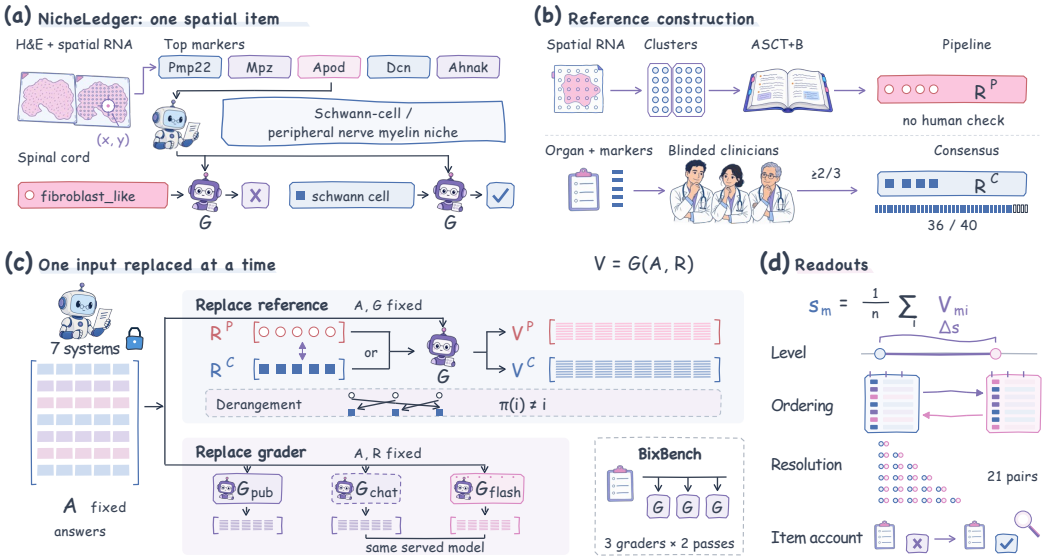


Figure 1: **A benchmark score has three inputs. Each experiment replaces one while holding answers fixed.** (a) One niche-naming item, its markers, a frontier answer, and the two references it is graded against. (b) Reference sources, with the clustering pipeline above and blinded clinicians below. (c) The two replacements, with the derangement control and the external benchmark. (d) The four readouts. Tiles, matrices and ranking marks are illustrative, not measured.

against another item's reference gives at most 8.3% (Section 4). Two physicians grading the same answers by hand give 30.6% and 86.1%. On seven of the 18 items gained, no clinician named the reference's cell type. Clusters marked by *Mpz* and *Pmp22* carry an automated fibroblast label, but all three clinicians call them Schwann cells. Only one of 21 system pairs changes order, while a paired test of the change in each gap, corrected for the 21 comparisons, finds the frontier's lead over four other systems widened. Replacing the grader moves forty fixed answers between 30.0%, 47.5% and 20.0% and changes the outcome of each of our three claims under at least one grader state (Section 5). BixBench releases answers that allow the same replacements, and the published ordering holds under three graders while the level moves by up to 4.5 points. Model-proposed, source-checked references change agent scores by at most 2.8 points without reordering them.

Stable ordering does not establish stable levels, gaps or test outcomes. Clinician consensus is not biological ground truth, and these effect sizes are benchmark-specific. Comparisons that held here need not hold under another grader. Our contributions are as follows:

- We replace the reference with answers and grader fixed (Section 4, Table 1). The reference alone moves a frontier model by 50 points on niche naming. A derangement control tests item correspondence, hand grading by two physicians independently supports the gain, and a descriptive classification and marker registry locate it.
- We replace the grader (Section 5, Table 2), testing the paper's comparative claims on identical answers under three grader states, two of them requested aliases that report the same served-model identifier, and we repeat both replacements on BixBench's released answers.
- We release the NicheLedger audit materials (Section 3), which pair H&E histology with spatial expression under a slide-disjoint split and include both references, the blind labels and checks that need no model call. The filtered release has distinct naming, retrieval-control and spatial uses (Appendix E), and every main-text number links to a command or archive (Appendix P).

2 Related Work

References that move scores. Construction artifacts can make a benchmark solvable without its intended reasoning. A hypothesis-only classifier reaches 67% on SNLI (Gururangan et al., 2018), an instance of shortcut learning (Geirhos et al., 2020). Natural-language understanding benchmarks

can also leave too little headroom to show improvement (Bowman & Dahl, 2021). Northcutt et al. (2021) find label errors on at least 6% of the ImageNet validation set and show that, on corrected labels, a smaller model overtakes a larger one once such errors are slightly more prevalent. Jin et al. (2026) find the same for text-to-SQL leaderboards, while replacing a test set (Recht et al., 2019) or its labels (Beyer et al., 2020) lowers the level or shrinks reported progress. In spatial omics, the annotations used to evaluate clustering are error-prone (Sun et al., 2026), and language models naming cell types from markers diverge from manual annotation through granularity and synonymy (Crowley et al., 2025). Human-annotation studies also argue for evaluating opinion distributions rather than majority labels (Pavlick & Kwiatkowski, 2019; Nie et al., 2020). In an agentic benchmark graded by a language model, we replace references with answers and grader fixed, trace each changed item, and measure effects on paired comparisons. Where Northcutt et al. (2021) report reordering under more prevalent errors, we find ordering largely preserved while gaps change.

Graders that move scores. A strong language-model judge agrees with human preferences over 80% of the time (Zheng et al., 2023). A science-exam grader correlates with expert marking at 0.948 over whole exams (Dinh et al., 2024), but agreement in expert domains is lower (Szymanski et al., 2025). Hosted models drift between dated versions (Chen et al., 2024a), and endpoints change effective behavior while remaining healthy (Leshin et al., 2026). Black-box probes can detect such changes cheaply (Chauvin et al., 2026). Judge version changes shift severity beyond a permutation null (Sunkavalli, 2026). Li (2026) attributes drift to the system or the judge with a fixed human-labeled anchor set, and Chen et al. (2026) correct judge-based estimates with a small set of human labels. Those works monitor, attribute or calibrate score levels. Because the grader behind one name changed during this project, we test the paper's own comparative claims under each grader state.

Agentic science benchmarks and audits. Audits report insufficient tests in SWE-bench Verified (Zhu et al., 2025), critical issues in over 25.7% of tasks across 168 benchmarks (Wang et al., 2026) and scores for 24 benchmarks against 46 practices (Reuel et al., 2024), while agent harnesses can fail to beat simple baselines (Kapoor et al., 2025). In biology, BixBench grades nearly 300 open-answer questions (Mitchener et al., 2025). Phylo revised a human-reviewed BixBench subset, BenchGuard audits task artifacts with frontier models, and an audit of three capsules traces most scored failures to keys and question wording (Phylo, 2026; Tu et al., 2026; Dai, 2026). LAB-Bench poses over 2,400 multiple-choice questions (Laurent et al., 2024), DiscoveryBench pairs real and synthetic discovery tasks (Majumder et al., 2025), GenoTEX evaluates gene-expression analysis agents (Liu et al., 2024), and SpatialBench scores 146 problems drawn from spatial-analysis workflows with deterministic graders (Workman et al., 2025). Language-model agents now run end-to-end single-cell and spatial transcriptomics analyses (Xiao et al., 2026; Lin et al., 2025; Niyakan & Qian, 2025), describe gene sets while verifying against domain databases (Wang et al., 2025c), and assign cell types after supervised and reinforcement learning (Fang et al., 2026). The evaluated agent arms come from a scaffold built for process-reward training (Lightman et al., 2024; Wang et al., 2024; Shao et al., 2024) (Appendix B).

3 NICHELEDGER AND THE REPLACEMENT DESIGN

Tasks and references. NicheLedger is built from public HEST-1k (Jaume et al., 2024) slides that pair H&E with spatial expression (Appendix C.2). Three families ask a model to name the tissue niche at a queried coordinate (niche naming, T1), to rank the top five marker genes of a named niche (biomarker ranking, T2), or to name the niches bordering a given niche with their genes (spatial relation, T3). Each reference is derived by per-slide Leiden clustering (Traag et al., 2019; Wolf et al., 2018) of expression and describes a cluster, its marker genes and its spatial neighbors. Clusters are named by marker overlap with the ASCT+B tables of the HuBMAP Human Reference Atlas (Börner et al., 2021), or by the organ and top marker gene when no cell type matches. No original reference was written or checked by a person. The queried coordinate is the mean position of the cluster's spots, and on 11 of the 40 niche-naming items of the evaluation set it lies more than 1.5 spot spacings from any spot, 10 of them among the 17 items on the older 200 µm array, because the cluster spans several tissue sections (Figure 4). Both references describe the cluster, not the queried location. Replacing its name does not repair the coordinate, where an image-reading system sees background. Train and test share no slides (Table C.1, Appendix C).

Blind labels. Three clinicians and three graduate students independently named the 40 niche-naming clusters of the evaluation set from the organ, the cancer type where applicable, and the cluster's top five markers, with the reference, query and image withheld. The clinicians agree with one another at chance-corrected $\kappa = 0.743$ and with the reference at 0.128 (Appendix D).

Systems and grading. We evaluate on 120 questions of the train-disjoint test split, the first 40 per family by identifier, because the full split of 174 is unbalanced (65 T1, 52 T2, 57 T3). A larger extended pool overlaps training slides and is used only for the untrained ceiling (Appendix C.1). The floor is a Qwen3.5-9B policy (Qwen Team, 2026) warm-started by chain-of-thought distillation (Appendix B.1), run inside an eight-node agent loop and, as a control, in a single pass. The ceiling is a frontier model, GPT-5.5 (OpenAI, 2026b), answering in one pass on the same tool evidence, with the reference never in the prompt (Appendix H). A variant trained with group-relative policy optimization (GRPO) (Shao et al., 2024) on data from before a pipeline correction is reported but not interpreted. The single-pass control runs the fine-tuned and base models, each with and without thinking, which makes seven systems in all (Appendix O). Answers are graded by a strict hosted DeepSeek judge that must return an explicit verdict (Appendix M). Niche naming uses the forward call alone. On the 45 sampled pairs to which two physicians both gave the same verdict rather than calling them unsure, the forward call agrees with them at $\kappa = 0.951$ where the bidirectional rule used in the benchmark's original floor-to-ceiling comparison reaches 0.687 (Appendices G and B). This is conditional agreement on decidable, agreed-upon pairs, not validation on all answers, systems or grader states.

The replacement design. Each experiment fixes the system answers and replaces exactly one other input, so a score change is attributable to that input. A rule fixed before any scoring gives an item a clinician-consensus reference when at least two of the three clinicians agree under the agreement study's semantic matcher. The reference takes the wording of the clinician in the most agreeing pairs, and items with no agreeing pair are excluded. Coverage is 36 of 40. For the grader, identical answers are regraded with identical code under three grader states, the model the legacy name returned for the published figures and the two current names (Section 5.1). We distinguish absolute success (level), system rank (ordering) and pairs a paired test separates (resolution).

Usable release. Consensus naming supports a cluster-label reference audit, not spatial ground truth. The biomarker family is a retrieval control. Filtered spatial items expose separate reference components, but stored predictions cover 15 of the 22 spatial items in the repaired release and establish no system separation. Appendix E gives the construction rules, coverage, component results and independent hand grading.

Four checks on the instrument. Four checks accompany the replacements to test whether the instrument could have produced a reported number (Appendix A). D1 evaluates answers written by domain experts using the benchmark's own grader. D2 tests, without model calls, whether each reference contains what its question asks for. D3 scores each component of a compound reference as well as the whole. D4 reruns a policy on the evidence its own loop gathered, with the loop removed.

4 Replacing the reference

4.1 Clinicians and the frontier recover the original reference at similar rates

Under the benchmark's grader and original references, three clinicians reach 30.0%, 27.5% and 30.0%, the frontier model 30.0%, the fine-tuned policy in a single pass 17.5%, three trainees 11.7% pooled, and that policy in the agent loop 2.5% (Figure 2). No paired exact test separates a clinician from the frontier ($p = 1.0$ against each, 0.39 against their union), which at 40 items is a statement about resolution rather than equality. The agent arm is strictly dominated by every clinician, winning no item any of them missed and losing eleven, ten and eleven ($p = 0.001$, 0.002 and 0.001, failure modes in Appendix N). At least one clinician recovers the reference on 16 of 40 items and all three on 7. Fallback strings alone do not explain this, since 8 of 19 of them and 10 of 21 real cell-type names are credited to no rater (Fisher exact $p = 0.76$).

The comparison depends on the grading rule. Under the label-only matcher of the agreement study the clinicians' union beats the frontier 20.0% to 5.0% ($b = 6$, $c = 0$, $p = 0.031$), but that matcher

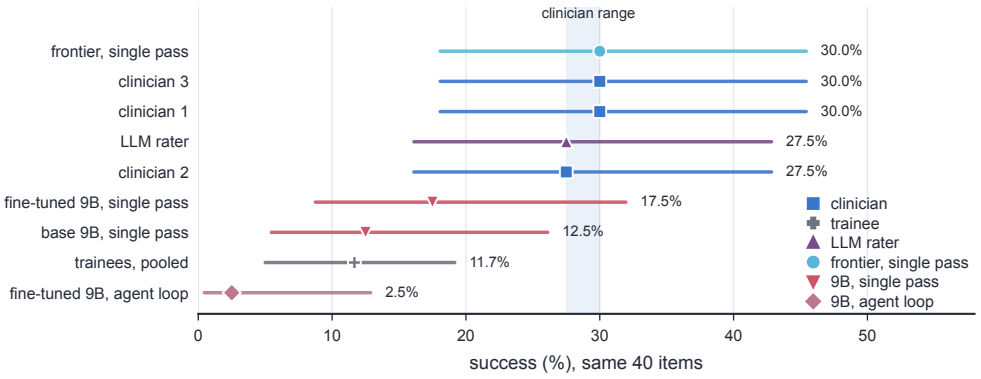


Figure 2: **The frontier falls within the clinician range under the original grader and references.** Wilson 95% intervals on the same 40 niche-naming items, and for the pooled trainees an interval that resamples items. The frontier reaches the top of the clinician range. No paired test separates it from any of the three. The LLM rater is Gemini 3.6 Flash on the clinicians' blind sheet (Appendix D). The published agent arm reaches 2.5%, dominated by every clinician and its own single-pass control.

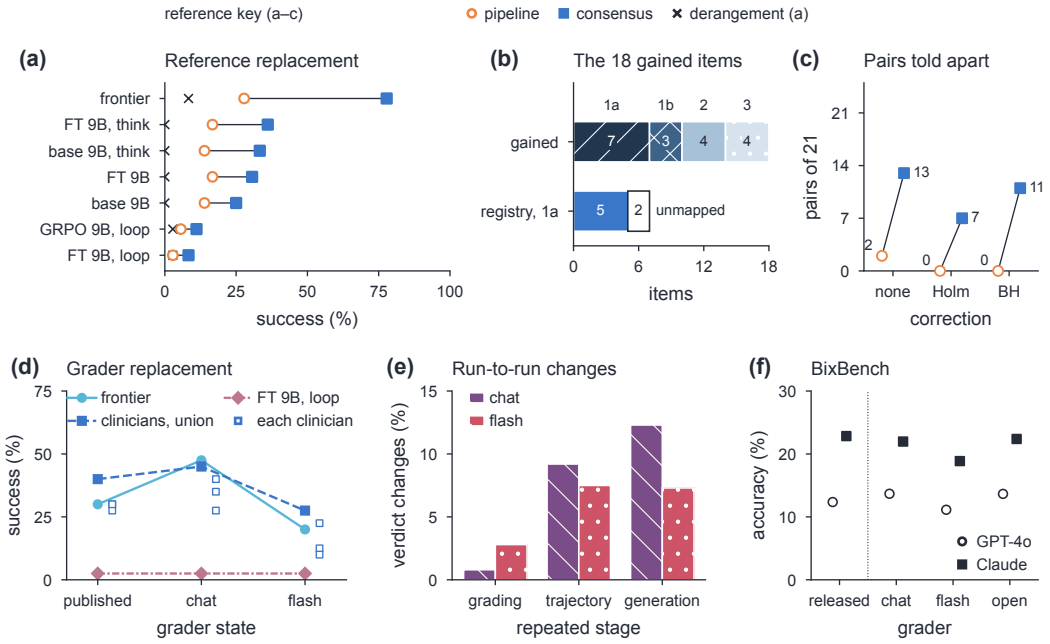


Figure 3: **Both replacements move levels and test outcomes while most orderings hold.** (a) Seven systems' fixed answers on 36 niche-naming items under one grader. (b) The 18 gains by class (Section 4.3) and registry mapping of the seven class-1a items. (c) Pairs an exact paired test separates at 0.05, uncorrected, Holm-adjusted and Benjamini–Hochberg-adjusted. (d) The frontier's 40 niche-naming answers under three grader states. (e) Verdicts that change across three runs when one stage is repeated. (f) BixBench's released answers under self-grading and three graders, two-pass means.

agrees with the two physicians' adjudication at $\kappa = 0.466$ against 0.951 for the forward judge and rejects 13 pairs they accept (Appendix G). This is reference recovery, not matched spatial capability.

4.2 Replacing the reference moves the frontier's level by 50 points

Re-graded on the 36 covered items in one pass, the frontier moves from 27.8% to 77.8% [61.9, 88.3] (paired McNemar $b=0$, $c=18$, $p < 0.001$), the fine-tuned policy in a single pass with thinking from

Table 1: **Every arm scores higher against the clinician consensus than against the original reference.** The 36 covered niche-naming items are graded in one pass, so only the reference varies. Derangement pairs each arm's answers with another item's consensus, preserving vocabulary but removing correspondence. Only the frontier's gain survives a Holm correction across the arms. One of the 21 arm pairs reverses between references, and neither reference separates that pair.

arm	original (%)	consensus (%)	paired b/c, p	derangement (%)
frontier, single pass	27.8	77.8	0 / 18, <0.001	8.3
fine-tuned 9B, single pass, thinking	16.7	36.1	0 / 7, 0.016	0.0
base 9B, single pass, thinking	13.9	33.3	0 / 7, 0.016	0.0
fine-tuned 9B, single pass	16.7	30.6	0 / 5, 0.062	0.0
base 9B, single pass	13.9	25.0	0 / 4, 0.125	0.0
GRPO 9B, agent loop	5.6	11.1	0 / 2, 0.5	2.8
fine-tuned 9B, agent loop	2.8	8.3	0 / 2, 0.5	2.8

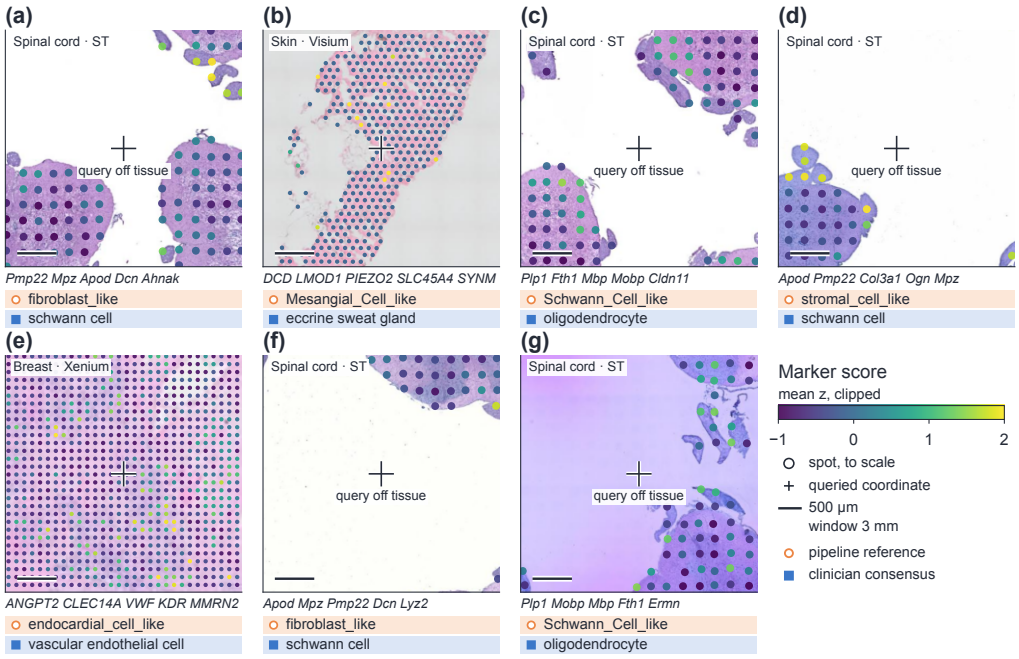


Figure 4: **The seven original references name cell types no clinician named.** All class-1a items, in identifier order. Each 3 mm H&E window is centered on the queried coordinate (cross), every spot drawn to scale and colored by the mean z-score of the five markers shown to clinicians. Corner tags give organ and assay, and strips give markers and both references. On five of the seven items the queried coordinate, a cluster's mean position, lies between tissue sections.

16.7% to 36.1%, and the agent floor from 2.8% to 8.3% (Table 1, Figure 3a). Consensus references replace many clustering strings with familiar cell types. The derangement control tests whether familiar wording alone earns credit, with four of seven arms at zero and none above 8.3%. This single derangement does not estimate permutation variability or validate each consensus name. Two physicians grading the 72 shuffled pairs by hand, not told which reference each was, jointly credit 11 and 31 of 36 ($b=0$, $c=20$, $p<0.001$), independently supporting the reference effect, not spatial accuracy or freedom from wording preferences (Appendix E). The clinician tie-break also matters. On six of the frontier's eight misses under the consensus it named what one clinician named, so the permissive three-clinician union raises it to 94.4%, not the adopted score (Appendix E). All figures concern covered items. Items on which no two clinicians agreed remain unscored.

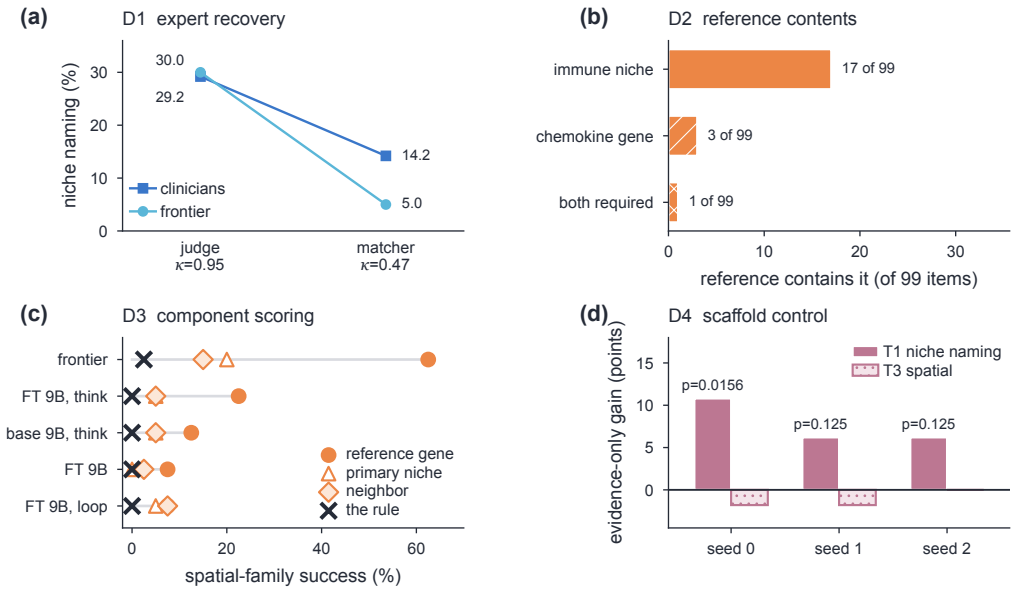


Figure 5: **Each check finds a way the instrument could have produced a number.** (a) D1: the judge scores the frontier level with the clinicians, the matcher below them. Each κ measures grader agreement with the clinicians' blind adjudication. (b) D2: of 99 questions, 17 references name an immune neighbor, 3 a chemokine and one both. (c) D3: conjunctive grading scores four systems at zero although their component scores separate them. (d) D4: keeping only the evidence line gains 10.8, 6.2 and 6.2 points on niche naming over three re-runs, and no gain on the spatial family.

4.3 The eighteen gained items

Reference replacement can correct an error or change the target. We classified the 18 gained items using a rule fixed before the table was built, but after inspecting cases. This is descriptive, not outcome-blind. It uses preexisting clinician labels, stored grader verdicts, the agreement matcher and two physicians' equivalence adjudications (Appendix F). On seven items (class 1a) no record links any clinician's label to the original reference, which names another cell type, for instance a fibroblast reference on clusters the three clinicians each call Schwann cells from *Mpz* and *Pmp22* (Figure 4a and f). Three original references are clustering strings that name no niche (class 1b). Four gains are wording or granularity (class 2), such as motor neuron against neuron, and four cannot be decided from the records (class 3).

A registry independent of the clinicians then tests the seven. PanglaoDB (Franzén et al., 2019), retrieved through OmniPath (Türei et al., 2021), maps both references to its cell types on five of the seven items. In a comparison fixed before any count, three of the cluster's top five markers belong to the clinicians' type on every one of the five, against at most one for the original reference's type. The registry does not favor consensus everywhere, and on the items that did not flip it favors the consensus on three of nine comparable items and ties on the other six. On ten of the 18 gained items, 27.8 of the 50 points, the original reference therefore names a cell type no clinician named or no niche at all. The other eight items, 22.2 points, are a change of target or undecided.

4.4 Orderings largely hold while the frontier's lead widens

One of the 21 arm pairs reverses between the two references (Spearman 0.91), and neither reference separates that pair. An exact paired test separates 2 pairs at uncorrected $p < 0.05$ under the original reference and 13 under the consensus, none and seven after Holm's correction, and none and 11 under Benjamini–Hochberg (Figure 3c). These tests are run per reference. A post hoc exact sign test on items gained by only one arm measures each gap's change directly. After Holm's correction the gap widens for four pairs, the frontier against both agent-loop arms and both single-pass arms without thinking, by 36.1 to 44.4 points, and for no pair without the frontier even uncorrected (Appendix F).

Table 2: **Each claim changes its outcome under at least one grader.** Claims use the 80 niche-naming and spatial items, and D3 the 40 spatial ones. *Published* is the model the legacy name returned on 2026-09-07, and *chat* and *flash* are the current names. Bold marks claims that fail under that grader. Domination is strict, so one item the floor wins and the ceiling loses (b) breaks it.

claim	published	chat	flash
D4, the agent loop costs	$+7.5$, p=0.031	$+8.8$, p=0.039	$+3.8$, p=0.25
D3, four systems at exactly zero	4 of 4	**0 of 4**	4 of 4
floor-to-ceiling domination	b=0, p=0.001	b=1, p<0.001	b=1, p=0.125

4.5 Spatial references that often cannot be matched

The spatial family fails differently. Across the splits, 99 distinct questions ask which immune niches border a region and which chemokines recruit into it. A model-free keyword rule finds an immune neighbor in 17 of their references, a chemokine in 3 and both in one (Figure 5b). Two physicians judged none of 20 sampled references to answer its question (Appendix G). By that rule, most of these items grade answers against references lacking what was asked. Questions were templated over niche names while references came from clustering output, and nothing checked the two against each other. Grading the compound reference as one verdict masks differences between systems (D3). Four systems whose reference-gene hit rates are 22.5%, 12.5%, 7.5% and 7.5% all score zero, and the frontier scores 2.5% while mentioning a reference gene on 62.5% of items (Figure 5c, Appendix A).

5 Replacing the grader

5.1 A grader name is not an identifier

Between the grading of 2026-09-07 and re-grading three days later, the model behind the requested grader name changed. Requests succeed without warning, and the served-model field, recorded since the re-grading, reports a different model than the one asked for. Identical-code re-grading of fixed predictions moved the frontier ten points on niche naming and fifteen on spatial items, and the two agent-loop arms one and two items. By the vendor's notice, the legacy name and the fast model's own name now both reach its fast model, yet their grading pipelines differ. On the frontier's forty niche-naming answers, with the same prompt, the two names' grading pipelines credit 47.5% and 20.0%, eleven items apart, while within-name re-grading changes at most 2.8% of verdicts (Section 5.3). In pooled diagnostic calls, the second name returns a verdict line on 53% of replies and the first on all, and a reply without one scores as incorrect. Only run-level parse counts were kept, so parsing and semantics cannot be separated. Both names report the same served model on every call, so that field cannot tell the two graders apart. Nor does it prove identical weights, and every version-pinned name probed is rejected (Appendix G). Pinned open-weight 9B, 14B and 32B graders, quantized or not, fail prespecified criteria. The unquantized 32B agrees with the clinicians at $\kappa = 0.86$ and 0.61 in two runs that differ in hardware and serving version, against the hosted grader's 0.95.

5.2 Three grader states change levels and comparative claims

On the forty niche-naming answers of Section 4.1 the frontier scores 30.0%, 47.5% and 20.0% under the three grader states (Figure 3d), and each of the three claims in Table 2 changes outcome under at least one, though the loop's penalty keeps its sign. A single pass on the loop's evidence tests this scaffold's text context (D4, Figure 5d), not agent loops generally. Its critic returns a constant score and its executor receives embedding dimensions, not embeddings (Appendix I). Under the published grader the loop wins none of the 80 items and loses six, all niche naming (exact McNemar $p = 0.031$), a penalty of 7.5, 8.8 and 3.8 points across the three states, not significant under the third (Appendices I and O). With the scaffold controlled, floor and ceiling score 8.8% and 15.0% on these two families, against 1.3% and 13.8% with the agent loop (Tables O.1 and B.1). These unadjusted claim tests are not cross-grader interaction tests (Table 2).

Across all states, the clinicians' union (Figure 3d) never separates from the frontier ($p = 0.39$, 1.0, 0.375) and wins items the agent arm misses, losing none ($p = 0.0001$, < 0.0001, 0.002).

5.3 Run-to-run changes and the limits of aggregation

Across three repeats per stage under the two current names, re-grading identical text moves 0.8% and 2.8% of verdicts, re-running the agent loop from a fresh seed 9.2% and 7.5%, and regenerating single-pass answers at temperature zero 12.3% and 7.3% (Figure 3e, Table G.1). These are observed rates from different arm pools, not variance components with intervals. Six graders from two vendors err in the same direction on the selected human-agreement sample, rejecting equivalents the clinicians accept but crediting none they reject. Majority voting reaches κ 0.69 to 0.81 against 0.95 for the best single grader (Appendix G).

5.4 Both replacements on an external benchmark

BixBench's GPT-4o and Claude 3.5 Sonnet agents are self-graded. The ordering holds under all three shared graders in each of two passes (every paired $p < 10^{-4}$). On released nonempty answers, the 10.5-point self-grading gap becomes 7.3–8.8, and Claude's accuracy moves by up to 4.5 points from released scoring and by 4.1 among our graders (Figure 3f, Appendix K, including the all-attempt denominator). Two aliases reporting matching identifiers disagree on 82 and 106 items, 63 and 78 with both replies parsed. Model-free checks find answer leakage of 2.4% on BixBench and 0.7% over eight LAB-Bench subsets against 60% in our spatial family (Appendix J).

For 165 questions in 29 capsules with source archives of at most 20 MB, GPT-6 Astra (OpenAI, 2026a) at Max effort proposed references from executed notebooks, blind to released references and answers, then checked them against the source. It and Claude Opus 5.5 (Anthropic, 2026), also at Max effort, reviewed blind the proposals not validated as equivalent. Agent replicas score 17% to 25% (Appendix L). Using 123 source-supported proposals and released references elsewhere moves no agent more than 2.8 points and narrows Claude's lead by 0.1 to 2.6 points under the two graders, all intervals including zero. Adding the 19 unvalidated proposals narrows it by up to 3.6 points, with uncorrected pointwise intervals excluding zero on image runs. Wording-only changes move scores by at most 0.8 points, mismatched references lower them by 14 to 19, and graders register 65% to 76% of substantive anchor changes. On 39 substantively different proposals, two blinded physicians both prefer 24 proposals and neither accepts the released reference on 30. Both accept both answers on four equivalent controls. Those 24 replacements move agent scores by at most 1.4 points and narrow Claude's lead by 1.0 to 1.3, all intervals including zero. Graders register 92% to 96% of the changes. Proposals are longer and show their derivations, which may sway reviewers, and source checks inherit unstated notebook choices, as on the one answer Verified-50 reviewers corrected (Phylo, 2026). This transfers the audit procedure, not the large effect or defect prevalence.

6 Limitations

Consensus labels do not validate spatial biology. Image necessity beyond expression remains unshown. T2 is a retrieval control, and T3 lacks spatial validation (Section 3). Six of the seven systems derive from one 9B model. External references are same-family-checked model proposals on a size-selected subset. The reference evidence rests on forty items and three clinicians, the hosted grading on one vendor, and all of it on our benchmark and one external one. Selected samples do not establish population-wide error rates. The four checks are not a certificate.

7 Conclusion

On fixed answers, clinician consensus moved the frontier's niche-naming score by 50 points. Derangement supports item correspondence, not biological truth, and on ten of the eighteen gained items the original reference named a cell type no clinician named or no niche at all. Changing the grader moved the same answers by 27.5 points and changed the outcome of three claims. Stable ordering did not imply stable levels, gaps or test outcomes. Releasing per-item verdicts and requested and served grader identifiers lets readers test these separately. Neither the effect sizes nor the comparisons should be assumed to generalize.

AI USE STATEMENT

We used generative AI tools to help implement the methods and write and edit the software code, to edit the text of this paper for readability, and to generate the icons in Figure 1. They also helped propose and refine hypotheses, design and give feedback on the methodology and experiments, clean and reformat the datasets, interpret results and translate the blind sheets, and Qwen3.5-9B generated the synthetic demonstrations that warm-started the fine-tuned arm. For BixBench, GPT-6 Astra proposed reference answers, and physicians reviewed the substantive changes blind (Section 5.4). The language models that appear in the experiments are objects of study. We reviewed all AI-assisted work and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses only previously published, publicly released data. HEST-1k redistributes spatial transcriptomics profiles and paired histology from prior studies whose consent and governance were established by the original investigators, and we introduce no new human-subject data, no identifiable patient information and no clinical records. We inherit the CC-BY-NC-SA 4.0 restriction that HEST-1k carries and state it on the released artifact.

The blind naming study asked six volunteer annotators to assign cell-type names to marker lists. No personal data was collected from them, the task carried no foreseeable risk, and their sheets are released with rater identity reduced to a panel label. Three physician researchers also volunteered, two to grade answer and reference pairs and all three to review BixBench references, under the same conditions, and their sheets are released with identity reduced to a reviewer label. These were expert annotation tasks on public research data, not studies of the annotators, and no institutional review or exemption was sought for them. We report the annotators' disagreement with our own reference labels, and we report an earlier coordinated labeling round as a control because coordination among raters can inflate their agreement.

Two risks are worth naming for downstream users. First, the T1 reference is derived from unsupervised clustering and, as this paper measures, diverges substantially from what independent clinicians name. Anyone using these labels as biological ground truth rather than as a benchmark target would be misusing them. Second, an agent that scores well here has not been validated for any clinical or diagnostic use, and the paper reports no wet-lab validation. We recommend that this benchmark be used to evaluate agents and not to support biological or clinical conclusions on its own.

REPRODUCIBILITY STATEMENT

The grader is a hosted model reached by name, and that name no longer returns the model that produced the published numbers. No version-pinned identifier exists to request instead, so the grading runs in this paper cannot be reproduced exactly by anyone, including us. This is the subject of Section 5.1 rather than a caveat to it. A reader can re-grade the released predictions under the graders reachable at the time and record which claims hold, as we did. The model-free checks can be recomputed without a grader. The clinicians' blind label analyses can be recomputed without new grader calls from the released sheets and stored semantic-matcher verdicts. Every graded comparison, including the two that held under all three grader states here, has to be re-established under the grader a reader can reach.

Within a run none of 720 repeated decisions changed, which does not show that the hosted service is deterministic. Two same-day runs on the same 400 rater-item pairs differed on 12 decisions (3.0%), and re-running the agent at temperature zero with a fixed seed does not reproduce earlier answers (Section 5.3, Appendix G).

The package recomputes every reported number from archived predictions, grades and annotation sheets. It does not retrain the policies, whose distillation demonstrations, training membership lists and adapter weights are not released, so the fine-tuned arms cannot be regenerated from it. Section 3 and Appendix C describe how every question is generated from public HEST-1k slides, how the reference answer is derived per slide, and how the slide-level and cross-tissue splits are constructed, with their overlap counts. Appendix P pins the software versions, random seeds and split definitions.

Appendix B.1 records the evaluated adapter configurations. Appendix M records the NicheLedger prompts and the BixBench grading wording, and the executable scorers preserve the literal strings on which the reported rates depend. Appendix D states the blind protocol, the matchers and the permutation and bootstrap procedures behind the human agreement numbers. Appendix H reports the compute actually used. The benchmark generation scripts, the evaluation code and the blind rater sheets are included in the supplementary material, which also contains the archived result files from which every table in this paper is computed. A single command checks every number in the main text from the released files and names for each the command that produces it or, where a hosted model, a GPU or another group's data would be needed, the archived result it is checked against. The model-free checks are also released as a standalone tool whose self-test recomputes this paper's spatial-family figures.

REFERENCES

Anthropic. Introducing Claude 3.5 Sonnet. Anthropic announcement, June 2024, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. Introducing Claude Opus 5.5. Anthropic announcement, September 2026, 2026. URL https://www.anthropic.com/claude-opus-5-5.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Katy Börner, Sarah A. Teichmann, Ellen M. Quardokus, James C. Gee, Kristen Browne, David Osumi-Sutherland, Bruce W. Herr, II, Andreas Bueckle, Hrishikesh Paul, Muzlifah Haniffa, et al. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nature Cell Biology*, 23(11):1117–1128, 2021. doi: 10.1038/s41556-021-00788-6.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.385.

Timothée Chauvin, Clément Lalanne, Erwan Le Merrer, Jean-Michel Loubes, François Taïani, and Gilles Tredan. Token-efficient change detection in LLM APIs. *arXiv preprint arXiv:2602.11083*, 2026.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024a. doi: 10.1162/99608f92.5317da47.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024b. doi: 10.1038/s41591-024-02857-3.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. *arXiv preprint arXiv:2601.05420*, 2026.

George Crowley, Tabula Sapiens Consortium, and Stephen R. Quake. Benchmarking cell type and gene set annotation by large language models with AnnDictionary. *Nature Communications*, 16: 9511, 2025. doi: 10.1038/s41467-025-64511-x.

Anqi Dai. What the score hides: an audit of BixBench — code and study. Zenodo, 2026. Version 1.0.1.

DeepSeek. Change log. DeepSeek API documentation, 2026a. URL https://api-docs.deepseek.com/updates/. Accessed September 17, 2026.

DeepSeek. DeepSeek V4 preview release. DeepSeek API documentation, 2026b. URL https://api-docs.deepseek.com/news/news260424/. Published April 24, 2026; accessed September 17, 2026.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. SciEx: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11592–11610, 2024. doi: 10.18653/v1/2024.emnlp-main.647.

Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Yifan Yang, Aidong Zhang, Jiawei Han, and Zhiyong Lu. Cell-o1: training LLMs to solve single-cell reasoning puzzles with reinforcement learning. *Bioinformatics*, 42(5):btag208, 2026. doi: 10.1093/bioinformatics/btag208.

Oscar Franzén, Li-Ming Gan, and Johan L. M. Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019:baz046, 2019. doi: 10.1093/database/baz046.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Google. Introducing Gemini 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber. Google blog, July 2026, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F.K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53798–53833. Curran Associates, Inc., 2024. doi: 10.52202/079017-1704. Datasets and Benchmarks Track.

Tengjun Jin, Yoojin Choi, Yuxuan Zhu, and Daniel Kang. Pervasive annotation errors break text-to-SQL benchmarks and leaderboards. *arXiv preprint arXiv:2601.08778*, 2026.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Jonah Leshin, Manish Shah, Ian Timmis, and Daniel Kang. Behavioral fingerprints for LLM endpoint stability and identity. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pp. 1327–1331, 2026. doi: 10.1145/3786335.3813194.

Yitao Li. Who drifted: the system or the judge? anytime-valid attribution in LLM evaluation pipelines. *arXiv preprint arXiv:2606.15474*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Zuwan Lin, Wenbo Wang, Arnau Marin-Llobet, Qiang Li, Samuel D. Pollock, Xin Sui, Almir Aljovic, Jaeyong Lee, Jongmin Baek, Ningyue Liang, Xinhe Zhang, Connie Kangni Wang, Jiahao Huang, Mai Liu, Zihan Gao, Hao Sheng, Jin Du, Stephen J. Lee, Brandon Wang, Yichun He, Jie Ding, Xiao Wang, Juan R. Alvarez-Dominguez, and Jia Liu. Spatial transcriptomics AI agent charts hpsc-pancreas maturation in vivo. *bioRxiv*, 2025. doi: 10.1101/2025.04.01.646731.

Haoyang Liu, Shuyu Chen, Ye Zhang, and Haohan Wang. GenoTEX: An LLM agent benchmark for automated gene expression data analysis. *arXiv preprint arXiv:2406.15341*, 2024.

Yang Liu, Lu Zhou, Xiawei Du, Ruikun He, Xuguang Zhang, Rongbo Shen, and Yixue Li. Benchmarking LLM-based agents for single-cell omics analysis. *Genome Biology*, 27:123, 2026. doi: 10.1186/s13059-026-03998-z.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Ludovico Mitchener, Jon M. Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096v1*, 2025. URL https://arxiv.org/abs/2503.00096v1.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9131–9143, 2020. doi: 10.18653/v1/2020.emnlp-main.734.

Seyednami Niyakan and Xiaoning Qian. PhenoGraph: A multi-agent framework for phenotype-driven discovery in spatial transcriptomics data augmented with knowledge graphs. *bioRxiv*, 2025. doi: 10.1101/2025.06.06.658341.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. GPT-6 Astra: A new generation of intelligence. OpenAI announcement, 2026a. URL https://openai.com/index/gpt-6-astra/.

OpenAI. Introducing GPT-5.5. OpenAI announcement, April 2026, 2026b. URL https://openai.com/index/introducing-gpt-5-5/.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019. doi: 10.1162/tacl_a_00293.

Phylo. BixBench-Verified-50. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/phylobio/BixBench-Verified-50.

Qwen Team. Qwen3.5-9B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. doi: 10.48550/arXiv.2408.00714.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21763–21813, 2024. doi: 10.52202/079017-0685. Datasets and Benchmarks Track.

Yusuf H. Roohani, Andrew H. Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300.

Jieran Sun, Kirti Biharie, Peiying Cai, Niklas Müller-Bötticher, Paul Kiessling, Meghan A. Turner, Søren Helweg Dam, Florian Heyl, Sarusan Kathirchelvan, Martin Emons, et al. Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering. *Nature Methods*, 23 (9):1814–1826, 2026. doi: 10.1038/s41592-026-03194-8.

Veerendra Kumar Sunkavalli. LLM judges as raters: A pre-registered audit of severity, halo, reliability, and version instability in LLM essay scoring on public corpora. *arXiv preprint arXiv:2608.29517*, 2026.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 952–966. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712091.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z.

Xinming Tu, Tianze Wang, Yingzhou Lu, Kexin Huang, Yuanhao Qu, and Sara Mostafavi. BenchGuard: Who guards the benchmarks? automated auditing of LLM agent benchmarks. *arXiv preprint arXiv:2604.24955*, 2026.

Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, Tamás Korcsmáros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021. doi: 10.15252/msb.20209923.

Chloe Xueqi Wang, Haotian Cui, Andrew Hanzhuo Zhang, Ronald Xie, Hani Goodarzi, and Bo Wang. scGPT-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. *bioRxiv*, 2025a. doi: 10.1101/2025.02.05.636714.

Hanchen Wang, Yichun He, Paula P. Coelho, Matthew Bucci, Abbas Nazir, Bob Chen, Linh T. Trinh, Serena Zhang, Ziyu Lu, Kexin Huang, Vineethkrishna Chandrasekar, et al. SpatialAgent: An autonomous AI agent for spatial biology. *bioRxiv*, 2025b. doi: 10.1101/2025.04.03.646459.

Junlin Wang, Federico Bianchi, Shang Zhu, Fan Nie, Yongchan Kwon, Bhuwan Dhingra, and James Zou. Automated benchmark auditing for AI agents and large language models. *arXiv preprint arXiv:2605.26079*, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: Self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025c. doi: 10.1038/s41592-025-02748-6.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Kenny Workman, Zhen Yang, Harihara Muralidharan, and Hannah Le. SpatialBench: Can agents analyze real-world spatial biology data? *arXiv preprint arXiv:2512.21907*, 2025. doi: 10.48550/arXiv.2512.21907.

Yihang Xiao, Jinyi Liu, Yan Zheng, Shaoqing Jiao, Jianye Hao, Xiaohan Xie, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Zhen Wang, Xuequn Shang, Zhijie Bao, Changxiao Yang, and Jiajie Peng. CellAgent: LLM-driven multi-agent framework for natural language-based single-cell analysis. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BsA2GNkJhz.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. Datasets and Benchmarks Track.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, et al. Establishing best practices in building rigorous agentic benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 184435–184475, 2025. doi: 10.52202/085713-5547. Datasets and Benchmarks Track.

APPENDIX

The appendix is ordered by the claim each part supports. Appendix A defines the four checks of Section 3, and Appendix B describes the evaluated agent, its weights and the original bracket. Appendices C–E document the benchmark, the blind human review of its niche-naming references and the repaired release. Appendices F and G support the reference and grader replacements of Sections 4 and 5, and Appendices J–L the analyses on benchmarks built by other groups. The remaining appendices document compute, a complete executor context, prompts, failure modes, the seven scaffold arms and reproduction details.

CONTENTS

A The four checks behind the measurements

B The evaluated agent and the original bracket

C NicheLedger construction, composition and data provenance

D Blind human review of the T1 reference

E The repaired release

F The eighteen items gained by replacing the reference

G Grader validation across protocols and states

H Compute resources

I A complete executor context

J Model-free checks on benchmarks built by other groups

K Replacing the grader on a benchmark we did not build

L External reference sensitivity on BixBench

M Full LLM prompts

N Failure-mode analysis

O The seven arms of D4

P Seeds, splits and software versions

A THE FOUR CHECKS BEHIND THE MEASUREMENTS

Each check asks whether the instrument could have produced a reported number, and each applies to any benchmark that grades free-text answers against references. None establishes soundness. These are minimum checks, not a certificate. Figure 5 in the main text shows one result of each.

D1, expert recovery (no model training, one grading run). Apply the benchmark's own grader, unchanged, to answers written by domain experts on the same items. If experts score near the models, the score may reflect reference agreement more than the ability the reference represents. This check alone cannot determine whether experts or systems answered the underlying task correctly.

D2, question–reference correspondence (no model calls). For each question, check whether the reference contains the kind of content the question asks for. Where it does not, a model that answers the question is graded against a reference lacking what was asked. The check reads released files, so a change of grader cannot move its numbers.

D3, component scoring (grader-mediated). For a compound reference, score each component as well as the whole. A conjunction of weak components is near zero by arithmetic, so a family score near zero can reflect the rule rather than the systems.

D4, scaffold control (grader-mediated on both sides, one inference pass per item). Re-run the same policy on the same evidence with the agent loop removed, holding prompt, sampling parameters and grader fixed. Report an un-fine-tuned model in a single pass as a second reference, because a scaffold that cannot beat it does not justify its complexity.

D1 in detail. Table A.1 shows four references that no clinician reproduces alongside the three clinicians' labels for the same clusters (Section 4.1).

D2 in detail. The spatial family comes from three templates sharing one rubric. Of the 57 test items, 23 ask for a niche with high spatial autocorrelation, 21 name the niche and ask for its neighborhoods,

Table A.1: **Some references are credited to no rater.** The D1 examples pair reference labels on the left with the three clinicians' labels on the right.

reference	clinicians C1 / C2 / C3
telocyte, interstitial-cell-of-Cajal-like	schwann cell / Schwann cell / meningeal fibroblast
fibroblast-like	Schwann cell / Schwann cell / Schwann cell
endocardial-cell-like	vascular endothelial / endothelial / vascular endothelial
colon adenocarcinoma, ARGLU1-high	intestinal epithelial / intestinal epithelial / REG3A+ epithelium

and 13 ask which immune niches border a named region and which chemokines drive recruitment. The rubric's first clause asks for a high-autocorrelation niche, which the second and third templates already supply, and 34 of the 57, or 60%, contain their own reference niche verbatim. We detect no scoring benefit from leakage, which is weaker than showing there is none. This paragraph uses the forward rule, which credits the frontier on 6 of the 40 scored items where the bidirectional rule credits one. Of the 40 items, 24 state their reference niche. Across seven systems the prediction repeats it on 46 of the 168 leaked item–system pairs, and the grader credits four, all the frontier's and all on the template where repeating the subject of the question is not optional. No system scores detectably better on leaked items than on clean ones. The frontier scores 16.7% on the 24 against 12.5% on the 16 (Fisher $p = 1.0$, which at these counts rules out neither direction), and the other six score zero on both. Instead, the leak makes the rubric's first clause vacuous on two of the three templates, so the conjunction rests on the remaining clauses and the primary-niche component tests anything only on the 16 items that do not give it away.

The third template fails a stronger check. Across the splits it has 99 distinct questions, and an immune niche appears among the reference neighbors in 17, a chemokine among the genes in 3, and both in one. The rule is fixed and ignores case. A question belongs to the template when it contains the phrase *immune niches border*. It is counted once across the four released splits by its slide, text and reference rather than its identifier, because the extended pool reissues sampled training questions under new identifiers. Counting by identifier would include 26 such copies and give 125 questions and 24 immune neighbors. A neighbor counts as immune when its name contains one of fifteen stems (immune, T cell, B cell, macrophage, plasma, myeloid, lymph, NK, dendritic, monocyte, neutrophil, mast, microglia, leukocyte, granulocyte), and a gene counts as a chemokine when its symbol starts with a CC, CXC, CX3C or XC ligand or receptor prefix followed by a digit. A keyword rule can miss a synonym, so these counts state what the references contain by that rule, and the two physicians' reading of 20 sampled references is the check on it (Appendix G).

The defect arises because questions use niche-name templates while references come from clustering output without checking correspondence between the two. The reference gene lists have a second problem. In each split, 17% to 24% of spatial items contain at least one mitochondrial, ribosomal, predicted-gene, or negative-control entry, the standard signature of technical artifact in this assay. A model is asked to reproduce these entries as biology.

D3 in detail. Spatial references combine a primary niche, its neighbors, and a five-gene list. The rubric grades them with one verdict. Table E.2 scores the components separately on the same 40 items. Gene matching is deterministic, so this component is not grader-mediated. The family is genuinely challenging, with the frontier naming the primary niche on 20% of items and a correct neighbor on 15%. Yet the grading rule assigns the same score to four systems whose reference-gene hit rates are 22.5%, 12.5%, 7.5% and 7.5%. It gives the frontier 2.5% although that model mentions a reference gene on 62.5% of items. A component hit does not show that the spatial relation was reasoned correctly, and the zeros partly follow from requiring every component at once. The exact zeros hold under two of the three graders (Table 2).

D4 in detail. The bracket's floor is a fine-tuned 9B policy inside an eight-node agent loop. Its ceiling is a frontier model answering in one pass on that loop's evidence. The comparison therefore varies policy and scaffold together. Running the same policy on the same evidence in a single pass separates policy from scaffold (Table O.1). Removing the loop raises niche naming from 2.5% to 17.5%. The executor context explains this contrast. It contains one line per completed node, of which one carries evidence. The others contain a plan identical across questions, a region count without

masks, two constant embedding dimensionalities, and an empty retrieval (Appendix I). Replaying that context with every line but the evidence deleted, so only the string varies, shows that the rest costs niche naming, 3.1% against 13.8% of 65 items, $b=0$, $c=7$, $p = 0.016$. No effect is detected on the spatial family, whose rates are too low to give the test power. Regenerating the arms moves 7.3% to 12.3% of verdicts (Table G.1), so D3 and D4 are best interpreted alongside these run-to-run rates rather than as exact point estimates.

B THE EVALUATED AGENT AND THE ORIGINAL BRACKET

The evaluated agent reads histology and spatial expression through tools, then emits an answer. Its final executor receives only text observations from earlier steps. The segmenter contributes a region count, the encoders contribute embedding dimensions, and the statistics tool provides cluster-level evidence. Masks and embedding vectors never reach the executor (Appendix I).

The benchmark was first built to train a step-level reward model for this agent from domain signals, namely the overlap of a predicted region with the reference niche, the spatial autocorrelation of marker expression, and rank agreement with the reference markers. In a probe of 99 fine-tuned-agent trajectories on training questions, these signals could be computed only for the 33 niche-naming trajectories. None was scored correct under the probe's strict-match outcome, so the step labels offered no contrast between steps leading to success and those that do not. No reward-trained policy was built after the benchmark was repaired, and none is evaluated in this paper.

B.1 WEIGHTS BEHIND THE EVALUATED AGENT ARMS

The agent arms use two LoRA adapters (Hu et al., 2022) trained on Qwen3.5-9B. The supervised fine-tuned (SFT) adapter has rank 16, alpha 32, and dropout 0.05 across all seven projection matrices. It was trained with learning rate 2×10^{-4}, cosine decay, a 3% warmup, two epochs, and an effective batch size of 16 using fused AdamW in bf16 with gradient checkpointing at seed 0. The archived GRPO adapter uses the same LoRA architecture. Its recorded configuration specifies learning rate 1×10^{-5}, linear decay, 200 optimization steps, 8 accumulation steps, four generations per prompt, $\beta = 0.04$, and a 448-token maximum completion at temperature 1.0 and seed 0. Neither adapter archive includes its initial distilled demonstrations or training question list. The GRPO arm predates the benchmark repair, and its training data cannot be recovered. We retain it as a caveated historical comparison, not evidence for a GRPO effect.

B.2 THE ORIGINAL BRACKET

These evaluations precede the replacements in Sections 4 and 5. The bracket in Table B.1 represents the original measurements, which are superseded by Table O.1 under the controlled evaluation. The arm marked with a dagger is a GRPO run whose training data predate a pipeline fix, retained for completeness and not interpreted. The frontier ceiling scores 37.5% here versus 38.3% in Table O.1, and T1 scores 25.0% versus 27.5%, because these measurements reflect two independent grading runs conducted one day apart. As shown in Appendix G, remote grader verdicts fluctuate on approximately 3% of evaluations across runs, corresponding to one question out of 120 on this arm.

Table B.1 reports the floor-to-ceiling bracket on the train-disjoint 120. The ceiling strictly dominates both floors. In paired McNemar tests, the floor never solves a question that the frontier ceiling misses ($b_{\text{floor}} = 0$), yielding $p \approx 0$ overall and $p = 0.002$ across the T2-excluded families. Absolute scores remain low. Excluding T2 to isolate tasks requiring biological reasoning rather than retrieval, the floor achieves 1.3% while the ceiling reaches 13.8%. T2 exhibits inflated scores across both policies (60.0% and 85.0%) because the query prompt exposes the reference gene panel, reducing the task to information retrieval. T1 niche naming peaks at 25.0% even for the frontier model, while T3 spatial relation remains near zero.

C NICHELEDGER CONSTRUCTION, COMPOSITION AND DATA PROVENANCE

Candidate generation. Three templates per task family yield 300 questions per family and 900 in all, populated from HEST-1k and HuBMAP HRA per-slide metadata. One template per family:

Table B.1: **The frontier ceiling strictly dominates the floor on the train-disjoint 120.** Entries are success rates (%) under the bidirectional rule, which rejects hypernym pairs that two clinicians accept (Appendix G). Paired McNemar tests give $b_{\text{floor}} = 0$, overall $p \approx 0$ and T2-excluded $p = 0.002$. *Overall*$^{-}$ excludes T2, the retrieval control whose evidence contains the reference genes. Overall columns show Wilson 95% intervals. †P0 is a caveated GRPO run from before the pipeline fix, with unrecoverable training data. Bold marks the best value per column.

Method	T1 (%)	T2 (%)	T3 (%)	Overall (%)	Overall^{-} (%)
SFT floor	2.5	60.0	0.0	20.8 (14.5, 28.9)	1.3 (0.2, 6.7)
P0 floor†	2.5	55.0	0.0	19.2 (13.1, 27.1)	1.3
GPT-5.5 ceiling	**25.0**	**85.0**	**2.5**	**37.5** (29.4, 46.4)	**13.8** (7.9, 23.0)

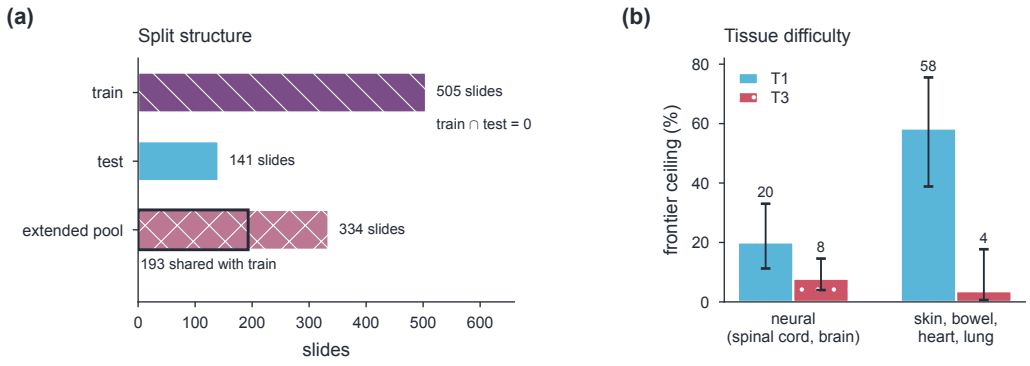


Figure C.1: **Only the untrained ceiling uses the training-overlapping extended pool.** (a) Train and test share no slides. The outlined overlap covers 193 of the extended pool's 334 slides, so this pool never evaluates a trained floor. (b) Frontier success with Wilson 95% intervals, recomputed from released splits and judged rows. Niche naming ranges from 20% on neural tissue to 58% on skin, bowel, heart and lung. Spatial-relation scores remain near the floor. No policy is tissue-specialized, so the contrast measures tissue difficulty, not train-to-test generalization.

- T1: "What tissue niche dominates the region centered at (c_y, c_x) in slide ⟨*slide*⟩?"
- T2: "Rank the top-5 marker genes for the ⟨*niche*⟩ niche in slide ⟨*slide*⟩."
- T3: "Which immune niches border the ⟨*primary niche*⟩ on slide ⟨*slide*⟩, and which chemokines might drive recruitment?"

No question was filtered by model agreement, so the released questions are the full candidate pool of 900. Slide-level partitioning allocates 646 questions to training, 80 to validation, and 174 to testing (Appendix P). The extended evaluation pool augments the 174 test items with 210 questions sampled from training using seed 42, accounting for the slide overlap with the training partition (Table C.1).

C.1 Benchmark composition

Table C.1 counts the questions and slides in each split, and Figure C.1 shows the split structure alongside the frontier ceiling by tissue group. The ceiling is computed on the extended pool after a quality filter removes 66 niche-naming items whose reference is a quality-control artifact or a marker-high fallback and one spatial item whose reference lists no neighboring niche. This leaves 317 questions. One group contains four organs (skin, bowel, heart and lung). The other contains the remaining slides, mostly spinal cord and brain. No slide belongs to both groups. Niche naming reaches 58.3% on the first group and 20.0% on the second, and the spatial family 3.6% and 7.8%. No evaluated policy is conditioned on tissue, so the contrast measures how difficulty differs between tissue groups, not generalization across them. Over all 317 questions the ceiling scores 40.4%, close to its 37.5% on the 120-question bracket of Appendix B.2.

Table C.1: **NicheLedger keeps test slides separate from training.** Counts are questions per split and family, with train $\cap$ test $= 0$ slides. †The extended pool contains the test split and overlaps training by 193 slides. Only the untrained frontier ceiling uses it, never the trained floor. Slides pair HEST-1k H&E with spatial transcriptomics.

Split	T1	T2	T3	Total	Slides
Train	204	220	222	646	505
Test (held-out)	65	52	57	174	141
Extended pool†	140	112	132	384	334

C.2 Dataset card

Name. NicheLedger v1.0.

Size. 646 training, 80 validation, 174 train-disjoint test, and 384 extended-pool questions across families T1, T2, and T3. The 99-trajectory probe of Appendix B is not part of the v1.0 question release.

Language. English.

Task types. Niche naming, biomarker ranking, spatial relation.

Source composition. All questions derive from HEST-1k slides. Each question carries its source slide identifier, platform and tissue, and the cluster record its reference was derived from. Per-family counts are in Table C.1.

Distribution statistics. The split is by slide, stratified by platform and tissue, and no slide appears in two of training, validation and test (Appendix P). The extended pool overlaps training slides and is used only for the untrained ceiling. The tissue breakdown of difficulty is in Figure C.1.

Known biases. (i) The 900 questions come from 711 slides, including 180 spinal cord, 119 brain, 83 breast and 54 bowel slides, so neural tissue dominates. (ii) 313 slides use the older spot array and 333 Visium, and Section 3 records that the older array is where queried coordinates most often fall between tissue sections. (iii) References are clustering outputs named against a marker vocabulary, not expert annotations, which is the subject of Section 4.

Licensing. Every question derives from HEST-1k, so the benchmark inherits its CC BY-NC-SA 4.0 terms. The code will be released under the Apache License 2.0.

Intended use. Research benchmark for scientific reasoning agents. Not suitable for clinical decision-making.

D Blind human review of the T1 reference

The T1 reference is produced by unsupervised clustering followed by automatic naming against a controlled vocabulary. We evaluated whether these labels match expert judgment. We presented the same 40 clusters, the T1 subset of the train-disjoint evaluation set, to two independent human panels and one language model. Every rater was provided the organ, the cancer type where applicable, and the cluster's top-5 marker genes, with the Leiden reference withheld, and recorded a single cell-type or niche name per row along with a confidence rating from 1 to 5. The first panel comprises three practicing physicians, while the second comprises three postgraduate researchers with formal training in biology. Raters completed their evaluations independently, without consultation, without shared documentation, and blinded to peer responses. We evaluate agreement among raters and with the reference using three protocols, exact normalized string matching, token-overlap fuzzy matching, and semantic matching with the same hosted grader used for benchmark scoring. We report chance-corrected Fleiss-style κ against a permutation null over 2,000 shuffles (Table D.1).

Contrasts below are paired item-level bootstraps over the 40 clusters with 10,000 resamples, so the two quantities being compared are always measured on the same instances. The 40 clusters come from 40 different slides, so resampling items is resampling slides. The raters are fixed, and every interval is conditional on these three clinicians and three graduate students rather than an estimate for their professions.

The clinicians agree with each other and not with the reference. Their agreement is $\kappa = 0.743$ with one another and $\kappa = 0.128$ with the Leiden reference, rising to 0.222 on the 21 clusters whose reference is a real cell-type name rather than a marker-high fallback. The paired gap in raw agreement, the item-level share of agreeing labels, is $+0.617$ (95% CI $[0.483, 0.742]$), and no bootstrap resample of 10,000 puts it at or below zero. Independent readers converge on an answer different from the clustering-derived reference. The mismatches are biologically grounded rather than arbitrary. On a *Plp1/Mbp/Mobp* signature the clinicians say oligodendrocyte while the reference says Schwann cell, and on a glycine-receptor signature they say inhibitory neuron while the reference says motor neuron.

Clinicians agree with each other far more than trainees do. The trainee panel agrees with itself at only $\kappa = 0.158$ and with the reference at 0.032. The paired gap between the two panels' internal agreement is $+0.542$ (95% CI $[0.408, 0.667]$), again with no resample at or below zero. The trainees also recognized that they were guessing. One recorded the lowest confidence level on 36 of 40 rows. Marker-only niche naming therefore separates these two panels, which a task at chance level would not. The panels differ in more than training, so the gap is not a measured effect of expertise, but it means the clinicians' divergence from the reference is not a restatement of task difficulty.

The language model has agreement rates close to those of the clinicians. Given the identical blind sheet, Gemini 3.6 Flash (Google, 2026), queried on 2026-08-01, agrees with the clinicians at $\kappa = 0.732$ against their 0.743 with each other, and reproduces the reference at $\kappa = 0.131$ against their 0.128. On the raw-agreement scale used above, its paired difference from the panel is -0.008 (95% CI $[-0.075, 0.058]$) on agreement with the clinicians and $+0.008$ (95% CI $[-0.067, 0.092]$) on agreement with the reference. Neither comparison separates the model from the panel, and both intervals bound the difference well inside the 0.617 gap that separates the panel from the reference. On these 40 marker-only items, the model's measured agreement is close to the clinician panel's, but the interval does not establish equivalence. The T1 ceiling against the clustering-derived reference mixes naming errors with reference disagreement. These data do not apportion the two contributions.

A control on the measurement. An earlier round of three student raters was coordinated by a group lead who circulated a shared sheet. Their apparent internal agreement, $\kappa = 0.842$, exceeds that of the independent clinicians. Chance-corrected inter-rater statistics assume independent rating, so coordination can inflate measured agreement. We report that round only as a control on the measurement and exclude it from every number above. Its advantage over the clinicians in agreement with the reference, 0.212 against 0.128 on all 40, shrinks to 0.248 against 0.222 on the clean 21, which suggests that it comes mostly from the marker-high fallback labels rather than from better naming.

We therefore score T1 semantically and treat its reference as a benchmark limitation. T2 contains a reference-panel retrieval shortcut, and the T3 reference has separate defects (Section 4.5).

E THE REPAIRED RELEASE

The release applies fixed rules to the train-disjoint test split (Table E.1), with no item edited by hand and no model called to build it. The original benchmark is released alongside it, because the audit cannot be checked without access to the original benchmark. The dataset card is part of the release. Filtering the niche-naming family instead of re-referencing it would drop the 26 references of the form organ, top gene, high. We evaluated and rejected that strategy because label format does not predict whether an annotator recovers the target. Among the evaluated items, 8 of 19 fallback references and 10 of 21 authentic cell-type names were credited to no rater under forward evaluation (Fisher exact $p = 0.76$). Applying heuristic filtering would have discarded valid items without resolving the underlying reference deficiency.

Human grading of the fixed answers. The language-model grader produced both scores, 27.8% and 77.8%, so a grader-induced gain would look the same. Two postdoctoral physician researchers,

Table D.1: **Independent readers agree with each other, and not with the Leiden T1 reference.** Chance-corrected κ on the same blind 40-cluster sample under two name matchers. For human panels, Within panel averages the three rater pairs. For the language model, it averages agreement with the three clinicians. The Gold columns report mean agreement with the reference over that panel's raters, on all 40 clusters and on the 21 whose reference is a real cell-type name. The coordinated panel shared a sheet and is a control on the measurement rather than a result.

		Within panel		Gold, all 40		Gold, clean 21	
Panel	n	Fuzzy	Sem.	Fuzzy	Sem.	Fuzzy	Sem.
Clinicians, independent	3	0.649	0.743	0.105	0.128	0.193	0.222
Trainees, independent	3	0.114	0.158	0.021	0.032	0.040	0.039
Language model (Gemini 3.6 Flash)	1	0.671	0.732	0.134	0.131	0.246	0.246
Coordinated panel, control	3	0.755	0.842	0.092	0.212	0.173	0.248

Table E.1: **The spatial filtering rules overlap.** Counts therefore do not sum to the 35 questions removed from that family. The table gives the repaired test composition by family after all three rules.

family	original	repaired	what happened
niche naming	65	36	clinician consensus, 25 never labeled, 4 without one
biomarker	52	52	kept, relabeled a positive control on the evidence pipeline
spatial	57	22	13 unanswerable template, 34 reference leaked, 10 artifact gene

who took no part in building either reference, therefore graded the frontier's 36 fixed answers against both references under a protocol and an analysis script fixed before any sheet was sent. Each of the 72 answer–reference pairs showed the organ, cancer type, cluster's top five markers, answer and one reference. Pairs were shuffled, with an item's two pairs never adjacent and the arm hidden. The question mirrors the grader's forward call, whether the answer names the same cell type or niche as the reference, allowing synonyms and reasonable differences in granularity. They worked independently, without AI tools or web search. Under the primary rule, which credits a pair when both say yes and counts unsure as not credited, they credit 11 of 36 answers against the original references and 31 against the consensus, all 20 changes in the same direction (exact McNemar $p < 0.001$), where the grader credits 10 and 28. Separately they credit 12 and 32, and 12 and 31, and crediting a pair when either says yes gives 13 and 32. They chose unsure on five and six of the 72 pairs. Dropping every item with an unsure pair leaves 11 of 30 and 30 of 30. They agree with each other at $\kappa = 0.913$ on whether a pair is credited, and with the grader's stored forward verdicts at 0.775 and 0.691. The arm was not shown, but clustering-derived strings often look automatically generated, so the grading is blind to the label and not to the style of a reference. This shows that the gain does not depend on the language-model grader. It does not show that each consensus name is right.

Residual headroom under the repaired reference. The consensus reference uses one clinician's wording, selected by a deterministic tie-break when at least two of the three agree. A prediction that matches another clinician's label, or the same label at another granularity, is therefore marked wrong because of the rule. Scoring every arm a second way, correct if the prediction is equivalent to any of the three clinicians' labels under the same grader and over every covered item rather than over the misses, the frontier reaches 94.4% [81.9, 98.5] against its 77.8%, while no other arm gains more than 5.6 points. Of the frontier model's eight remaining errors, six match an individual clinician's annotation. For example, the model predicted *REG3A-positive intestinal crypt epithelial niche* against the consensus reference *intestinal epithelial cell*, reproducing the exact phrasing of the third clinician. In another instance, the model predicted *neuronal gray-matter niche* against the consensus *excitatory neuron*, and on two items identified non-specific, ribosome-dominated profiles consistent with notes from the dissenting clinician. Two of the 36 items stay unmatched under this lenient reference, which bounds the tie-break's share of the eight errors without showing that the other six answers are biologically correct. The union of three labels is a lenient upper reference rather than a scoring rule. Three chances to match raise what a wrong answer can achieve by chance, and the derangement control was run for the consensus reference, not the union. The union is therefore not used to score the arms.

Table E.2: **Conjunctive scoring hides component differences between systems.** D3 scores the spatial family per component and as a conjunction. Four systems that the components separate all receive zero in the last two columns.

	Success (%)					
system	niche	$\geq$1 nbr	$\geq$1 gene	$\geq$3 genes	conj.	as graded
frontier model	20.0	15.0	62.5	32.5	2.5	2.5
fine-tuned 9B, 1 turn, thinking	5.0	5.0	22.5	7.5	0.0	0.0
base 9B, 1 turn, thinking	5.0	5.0	12.5	0.0	0.0	0.0
fine-tuned 9B, 1 turn, no thinking	0.0	2.5	7.5	0.0	0.0	0.0
fine-tuned 9B in the agent loop	5.0	7.5	7.5	0.0	0.0	0.0

Table E.3: **The repaired spatial subset does not separate system pairs.** Stored predictions cover 15 of 22 released items. No scoring rule separates any of the 21 arm pairs at $p < 0.05$. The ordering is plausible, but $n = 15$ cannot support a claim about it.

	Success (%)				
arm	conjunction	primary niche	any neighbor	$\geq$1 gene	$\geq$3 genes
frontier, single pass	6.7	13.3	26.7	20.0	20.0
sft 9B, single pass, thinking	0.0	13.3	13.3	6.7	0.0
base 9B, single pass, thinking	0.0	13.3	13.3	0.0	0.0
sft 9B, agent loop	0.0	13.3	20.0	0.0	0.0
GRPO 9B, agent loop	0.0	0.0	26.7	0.0	0.0
sft 9B, single pass	0.0	0.0	6.7	0.0	0.0
base 9B, single pass	0.0	0.0	6.7	0.0	0.0

Consensus construction and its control. An item receives a consensus reference when at least one of the three pairs of clinicians agrees under the agreement study's semantic matcher. The wording is that of the clinician who appears in the most agreeing pairs, and a tie goes to the clinician listed first in a fixed order. The derangement control pairs each covered item with another covered item. A permutation of the sorted item identifiers is drawn with seed 20260908 and redrawn until no item maps to itself, and every arm's answer is then graded against the paired item's consensus reference with the same prompt and protocol as the two references it controls. This single permutation shows that the answers do not match an arbitrary consensus name. It does not measure score variation across permutations or against a mismatched reference from the same organ.

Reproducibility of component scores. The component scoring was re-run on the original 40 spatial questions before being used on the repaired subset of Table E.3. Of 35 arm-by-component cells, two differ from Table E.2. Both are grader-mediated, the frontier's primary niche at 8 items against 9 and one arm's neighbor component at 2 against 3. All fifteen deterministic gene-matching cells reproduce exactly. This directly checks that the run-to-run component measured in Appendix G arises in the grader rather than the data path.

F THE EIGHTEEN ITEMS GAINED BY REPLACING THE REFERENCE

The rule, fixed before the table was built. The rule was written after four gained items had been read but before the rest. The classes therefore describe the gained items and are not an outcome-blind test. Gained items are those for which the fixed frontier answer is graded incorrect against the original reference but correct against the clinician consensus in the evaluation supporting Table 1. All classifications rely strictly on records archived prior to the analysis, including (a) independent clinician blind annotation sheets containing labels, confidence scores (1 to 5), and notes recorded without seeing the reference, (b) stored forward and reverse grader verdicts for each clinician label evaluated against the original reference, (c) stored decisions from the semantic agreement matcher, and (d) independent physician equivalence reviews conducted with references visible and grader verdicts hidden. A clinician label is *linked* to the reference when any of (b), (c) or (d) equates them. Taking the first rule that applies, an item is a wording or granularity change when a label of a clinician

Table F.1: **The 18 gained items fall into classes fixed before the table was built.** Identical pairs are merged. Class 1a names a cell type no record links to any clinician's label, 1b is a fallback string, 2 is wording or granularity, and 3 is undecided.

original reference	clinician consensus	items	class
`fibroblast_like`	schwann cell	2	1a
`stromal_cell_like`	schwann cell	1	1a
`Schwann_Cell_like`	oligodendrocyte	2	1a
`Mesangial_Cell_like`	eccrine sweat gland	1	1a
`endocardial_cell_like`	vascular endothelial cell	1	1a
`BONE_TPM1_high`	skeletal muscle cell	1	1b
`ColonAdenocarcinoma_ARGLU1_high`	intestinal epithelial cell	1	1b
`BRAIN_NRSN1_high`	GABAergic interneuron	1	1b
`motor_neuron_like`	neuron	2	2
`UTERUS_GRCH38_____SPARC_high`	fibroblast	1	2
`EYE_TCAP_high`	extraocular muscle cell	1	2
`InvasiveDuctalCarcinoma_RPL23_high`	fibroblast	1	3
`PROSTATE_RPS14_high`	smooth muscle cell	1	3
`InvasiveDuctalCarcinoma_ENSG00000269028_high`	low quality / unassigned	1	3
`Blood_Vessel_Endothelial_Cell_rich`	antigen-presenting cell	1	3

inside the consensus is linked. It is undecided when only the dissenting clinician's label is linked, when an adjudicator was unsure about a consensus pair or the two adjudicators disagreed, or when the consensus clinicians' median confidence is 2 or lower. Otherwise the records contradict the reference, which either names a cell type (class 1a) or is a fallback string of the form organ, top gene, high (class 1b). The stronger interpretation, that the reference caps the score, required at least ten of the 18 in class 1 under the rule fixed beforehand.

Result. Seven items fall in class 1a, three in 1b, four in class 2 and four in class 3 (Table F.1). Class 1 reaches ten, the threshold fixed beforehand, only by including the three fallback strings. Without them, the count is seven. The main text therefore reports the counts rather than describing the reference as capping the score. On the seven class-1a items the consensus clinicians' median confidence is 4 or 5, and on three of them both physicians, seeing the reference, judged the clinicians' label not equivalent to it. The result is exactly at the threshold, and one item could move in either direction. The endocardial reference is lineage-related to the clinicians' vascular endothelial cell although the tissue is breast, and one undecided item is undecided only because a single adjudicator marked it unsure.

An independent marker registry. PanglaoDB (Franzén et al., 2019), retrieved through OmniPath (Türei et al., 2021) on 2026-09-17, contained 8,211 marker records over 178 cell types. The comparison was fixed in writing before any count. Each reference name is mapped to registry cell types by tokens, dropping generic terms such as cell, like, and rich. Mapping proceeds by assigning an exact token-set match when available, falling back to registry types whose tokens form a proper subset of the query string, and marking entries as unmappable if no subset match exists. For each mappable name the score is how many of the cluster's top five markers the registry lists for that type. Both references map on five of the seven class-1a items, and on all five the consensus type scores 3 against at most 1 for the original type, the same with canonical markers only. Both controls behave as expected. The five items whose two references map to the same registry type all tie, and across the 16 comparable items the consensus type scores 2.25 markers on its own item against 0.218 on other items, so the count follows the item rather than the length of the registry's list. The registry does not favor the consensus everywhere. On the 18 items that did not flip, 9 are comparable and it favors the consensus on 3 of them, with the other 6 tied. The registry is independent of the clinicians and of the models, and it need not be independent of the marker vocabulary the original references were named against.

Separations under a multiplicity correction. The 21 arm pairs were each tested with the exact McNemar test of Table 1. Uncorrected, 2 pairs are separated at $p < 0.05$ under the original reference

Table F.2: **Replacing the reference widens the frontier's lead over the other arms.** Gaps are the frontier's success rate minus the other arm's, in points, on the 36 covered niche-naming items under one grading pass. The gains column counts items gained only by the frontier or only by the other arm. p is the exact two-sided sign test on those items. The six pairs with uncorrected $p < 0.05$ are shown.

frontier against	original	consensus	change	gains	p	Holm
GRPO 9B, agent loop	22.2	66.7	+44.4	16 / 0	3.1×10^{-5}	rejected
fine-tuned 9B, agent loop	25.0	69.4	+44.4	16 / 0	3.1×10^{-5}	rejected
base 9B, single pass	13.9	52.8	+38.9	14 / 0	1.2×10^{-4}	rejected
fine-tuned 9B, single pass	11.1	47.2	+36.1	14 / 1	9.8×10^{-4}	rejected
fine-tuned 9B, single pass, thinking	11.1	41.7	+30.6	12 / 1	0.0034	—
base 9B, single pass, thinking	13.9	44.4	+30.6	13 / 2	0.0074	—

Table G.1: **Three stages have different observed verdict-change rates.** Each column uses one grader on $n = 120$ items, with unregenerated arms as contemporaneous controls. Three runs per cell give observed rates, not variance components with intervals. The grading, trajectory and generation cells pool 360, 120 and 480 arm-item verdicts from three, one and four arms, so they describe each stage and do not compare the stages under matched arms. Grading changes least under both names. Generation changes most under chat and is level with trajectory under flash. Every cell compares three runs, counting a verdict as changed if it differs across them. Chat trajectory runs were graded together a week later than the rest. Under flash, 76% to 79% of replies carry a verdict. Missing verdicts count as incorrect.

		item verdicts changing (%)	
layer	what was re-run	chat-routed	flash-routed
grading	nothing; the same text graded again	0.8	2.8
trajectory	the agent loop re-planned, tool pipeline re-run	9.2	7.5
generation	single-pass answers regenerated, identical configuration	12.3	7.3

and 13 under the consensus. Under Holm's procedure at a family-wise error of 0.05 the counts are 0 and 7, and under Benjamini–Hochberg at a false discovery rate of 0.05 they are 0 and 11.

The change in each gap. Those counts come from two families of tests, one per reference, and the difference between them does not test whether any gap changed. For arms A and B and item i, let d_i be the change in $x_A(i) - x_B(i)$ when the consensus replaces the original reference, with x the forward verdict. Swapping the two references within an item flips the sign of d_i, so under the null that the replacement leaves the gap unchanged the exact two-sided sign-flip test on $\sum_i d_i$ is valid. No arm loses an item under the replacement, so every d_i lies in $\{-1, 0, 1\}$ and the test reduces to the exact sign test on the items one arm gains and the other does not. Table F.2 lists the six pairs with uncorrected $p < 0.05$, all of them the frontier against another arm. Holm's procedure at a family-wise error of 0.05 rejects four and Benjamini–Hochberg six. Among the 15 pairs without the frontier the smallest p is 0.0625, for four pairs whose gap widens by 13.9 points on five items only one arm gains. This test was not fixed in advance. The p-values of Table 2 are unadjusted. Crossing 0.05 under one grader but not another does not itself test whether an effect changed. We therefore test the change in each gap directly.

G GRADER VALIDATION ACROSS PROTOCOLS AND STATES

The main NicheLedger success rates use one instrument, a hosted DeepSeek grader requested by name with strict verdict-only output, under one of two protocols. The other graders of this appendix, the BixBench graders and the physicians' hand grading are described where they are reported. The bracket of Appendix B.2 asks in both reference-candidate orders and credits only when both return CORRECT, which is the published rule and a lower bound. Sections 4.1 and 4.2 use the single forward call, because the human adjudication below shows the bidirectional rule rejects hypernym pairs two clinicians accept. Section 5 replaces this instrument, so this appendix records which aspects have and have not been checked.

Table G.2: **Protocol mismatch inflates the human–machine gap.** On the same 40 items, bidirectional rather than forward-only grading costs people 5 to 10 points and systems 0 to 5, inflating their gap by about six. Rater rows come from the first of two same-day grading runs, which differs from the second, used in Section 4.1, on 12 of 400 decisions.

	forward only (%)	bidirectional (%)	change (points)
clinician C1	32.5	27.5	−5.0
clinician C2	30.0	20.0	−10.0
clinician C3	27.5	17.5	−10.0
language-model rater	32.5	22.5	−10.0
trainees, pooled	13.3	8.3	−5.0
frontier, single pass	30.0	27.5	−2.5
fine-tuned 9B, single pass	17.5	17.5	0.0
fine-tuned 9B, agent loop	2.5	2.5	0.0

The model reached through each name. Responses never report a model version, so we identify the grader by the name requested and by what the vendor documents behind it. The vendor's V4 announcement of 2026-04-24 routes the legacy chat name to the non-thinking mode of its V4 fast model and the legacy reasoning name to that model's thinking mode, and announces both names' retirement for 2026-07-24 (DeepSeek, 2026b). Both names kept answering after that date. By the vendor's account, then, every grading reported here before 2026-09-10 came from that fast model, and the reasoning-endpoint checks below came from its thinking mode. The vendor's changelog releases V4.1 Flash on 2026-09-10, the day of our re-grading, routes the older fast names to it, and keeps the pro model past 2026-09-14, the date it had been due to be retired (DeepSeek, 2026a). A probe of the API on 2026-09-17 lists two models, the fast and the pro. The legacy chat and reasoning names, the fast name and the older versioned fast name are all answered by the fast model, the pro name by itself, and every other versioned name tried is rejected with HTTP 400, including three spellings of the new fast model's version. A response names the served model but not its version or the routing that produced it.

Protocol symmetry. Human raters and systems have to be graded under the same rule. Grading raters with the forward call alone and systems bidirectionally builds in an asymmetry, which Table G.2 quantifies. The bidirectional rule is a conjunction over the forward call and strictly removes credits by construction. Over all ten rater sheets of the second run, the coordinated control included, it removes 20 of the raters' 100 forward credits, and over the seven systems 5 of their 38, rates of 20.0% and 13.2% that these data cannot separate (Fisher exact $p = 0.46$, descriptive only, since credits on one item are not independent). The raters lost more points because they held nearly three times as many forward credits. The reverse call penalizes granularity differences in either direction, as seen when evaluating `Ganglion_cell_like` against *neuron*, `basal_epithelia_like` against *keratinocyte*, or `stromal_cell_like` against *cancer-associated fibroblast*. Bidirectional grading therefore constrains granularity rather than testing biological validity alone.

Test-retest. Two arms were re-graded three times each over the same 120 questions at temperature zero. The frontier arm scores 46, 46, 46 and the published floor 25, 25, 25, and no item changes verdict in any of the 720 decisions. Between runs verdicts differ, even within one day. The same forward grading of the same 400 rater-item pairs, run twice on one day with identical code and identical inputs, disagrees on 12 of 400 decisions, spread over nine of the ten raters. The run-to-run component is not sampling noise inside a batch, and it need not be a change of model either, since a pinned open-weight grader with fixed weights at temperature zero differs on 13 to 35 of the same 400 between two back-to-back runs (below). Whatever the cause, the difference is roughly three decisions in a hundred. A comparison whose margin is one or two items in 120 is of the same order.

A known-answer probe. The diagnostic probe contains fourteen verification pairs with unambiguous ground-truth relationships derived from the benchmark vocabulary, comprising six identical cell types with alternative valid phrasings, six distinct biological lineages, and two degenerate candidates representative of agent formatting errors. The published grader is right 14 times out of 14 in both protocols. The same provider's reasoning endpoint, by its account the same fast model in thinking

Table G.3: **The pinned open-weight graders tested here fail the pre-specified checks.** κ measures agreement with the two clinicians' consensus on 45 pairs. Two values denote two runs with identical inputs. The hosted grader is the published reference, not a candidate.

grader	probe, forward	κ vs clinicians	errors, strict / lenient	changed between runs
hosted, published	14/14	0.951	1 / 0	12 of 400, same day
Qwen3.5-9B	6/14	not run	0 / 8, on the probe	—
Qwen3-14B	13/14	0.535, 0.500	11 / 0	26 of 400, back to back
Qwen3-32B, 4-bit	14/14	0.769, 0.857	5 / 0	13 of 400, back to back
Qwen3-32B, unquantized	14/14	0.857, 0.609	3 / 0	35 of 400, back to back

mode, is right 14 forward and 13 bidirectionally. A 9B model given the identical prompt is right 6 times and errs only in one direction, returning CORRECT on all eight pairs that are wrong, including the reference `adipocyte_like` against the candidate *unknown* and `fibroblast_like` against the raw cluster identifier `leiden_0`. Grader capacity is therefore not a free parameter of this benchmark. The probe requires 28 calls.

A pinned open-weight grader. A possible remedy is a self-hosted open-weight grader, whose weights cannot change during evaluation. We tested it against three criteria fixed before any call. A grader had to score 14 of 14 on the probe, reach $\kappa \geq 0.90$ against the clinicians' consensus on the 45 pairs they both decided and agreed on, and give identical verdicts across two runs with identical inputs. The 32B checkpoint (Qwen3-32B (Yang et al., 2025) in place of the planned Qwen2.5-32B, so that both sizes come from one model family) and a larger token limit were chosen after the first probe calls and before any agreement result existed. No grader we could serve passes (Table G.3). At 9B it credits wrong answers. At 14B and 32B, quantized or not, it errs the other way, rejecting equivalents the clinicians accept and never crediting a pair they reject. Both 32B graders pass the probe and still fail agreement with people, so a known-answer probe is necessary and not sufficient. Pinning the weights does not pin the verdicts. Two back-to-back runs at temperature zero on the same inputs differ on 13 of the 400 rater pairs in 4-bit and on 35 unquantized, and the unquantized grader's agreement with the clinicians falls from 0.86 in one run to 0.61 in the next. The unquantized runs used different hardware and a different serving version, so the two churn figures are not a clean test of quantization, but neither is zero. The hosted grader also fails the third criterion, changing 12 of the same 400 decisions within one day. Reproducibility therefore does not distinguish the open graders from it. Their distinguishing failure is agreement with people. We did not test any model larger than 32B.

Grader ensembles. The criteria were fixed before these results were computed. A consensus counts only if it reaches $\kappa \geq 0.951$, the best single grader, and changes no more than 13 of 400 verdicts between runs. Majority voting fails the agreement criterion in every pool (Table G.4). Crediting an item when any member credits it meets the agreement criterion in every pool. Its churn cannot be measured on the hosted and six-grader pools, whose hosted members were graded once, and on the open trio it is 23 of 400. We advise against adopting majority ensembles for three structural reasons. First, under one-directional error, an ensemble that credits any positive verdict cannot score below its strongest constituent, rendering the apparent improvement an artifact of aggregation. Second, the performance margin between the ensemble and the best individual grader represents only one item out of 45. Third, including a permissive 9B grader that accepts all eight negative probe controls causes the ensemble to inherit all false-positive decisions. Ensembling does not make the evaluation stable either. The open-weight trio, which constitutes the only fully replicated pool, exhibits verdict shifts on 16 of 400 evaluations under majority voting, which lies directly within the variation range of its constituent models (13 to 35).

Nested matching instruments. The naming-agreement study uses a second matcher that sees two labels but neither the question nor the family guidance. On the same 400 rater-item pairs it credits the clinicians at 14.2% where the judge credits them at 21.7% under the corrected protocol and 30.0% under the forward-only one. The two are not independent estimates of one quantity. Under the forward protocol the judge credits 48 items the matcher does not while the matcher credits 3 the

Table G.4: **Majority aggregation falls below the best single grader on clinician agreement.** κ uses the clinicians' 45-pair consensus, three rules and three pools. The best single grader scores 0.951. In two runs, the open trio, the only fully re-run pool, changes 16, 21 and 23 of 400 verdicts under the three rules, versus 13 to 35 for its members.

rule	hosted trio	open trio	all six
majority	0.812	0.812	0.687
unanimous	0.535	0.432	0.432
any member credits	1.000	0.951	1.000

Table G.5: **The forward grader matches clinician agreement on the decided pairs.** Both adjudicators decided and agreed on these 45 of 100 sampled items, so the agreement is conditional on them. The forward call's agreement with them equals the two physicians' agreement with each other on the 46 pairs both decided, which is a comparison and not a ceiling. The bidirectional rule falls short, and all seven of its disagreements are hypernym pairs.

protocol	κ vs human consensus	judge yes, humans no	judge no, humans yes
forward, one call	0.951 [0.835, 1.0]	0	1
bidirectional AND	0.687 [0.476, 0.867]	0	7
human vs human	0.951 [0.836, 1.0]	—	—

judge does not ($\kappa = 0.607$), and on the clinicians alone the nesting is exact, 19 against 0. Human adjudication is needed to assess which matcher is right.

A second grader. The same predictions re-graded through the same provider's reasoning endpoint, by its account the same fast model in thinking mode, under the identical prompt and rule, move no arm's overall rate by more than 0.9 points, with per-arm agreement κ from 0.875 to 1.000. Agreement is perfect on the retrieval family and lowest on niche naming, 0.60 to 0.79, which is the family whose reference the rest of this paper shows to be contested. On the human raters the second grader is systematically the more generous, moving C1 from 27.5% to 35.0% and C3 from 17.9% to 25.6%, so absolute levels move by 5 to 17 points with the grader configuration while the rankings shown here hold. Evaluated on items parsed by both configurations, clinician C1 attains 35.0% versus 32.5% for the frontier model ($p = 1.0$), C2 matches the frontier at 27.0% ($p = 1.0$), and C3 reaches 25.6% versus 30.8% ($p = 0.69$). Forty-three of 2,420 calls spent their budget on reasoning and emitted no verdict. Those decisions, 1.8% of the calls, are excluded rather than counted as incorrect, which is the one place in this paper that departs from that rule. This is a same-provider check on stability. A cross-family grader was not run, and the released code takes one as a parameter.

Validity, against two clinicians. One hundred of the grader's decisions were sampled, 50 it called correct and 50 it called incorrect, across rater strata. Two clinicians reviewed them with the machine verdict withheld and a different presentation order for each. Both answered all 100. They mark 45 and 52 as *unsure*, which itself measures uncertainty about the reference. On half the pairs a clinician will not say whether the candidate matches. On the 46 where both give a verdict they agree on 45, $\kappa = 0.951$ [0.836, 1.0], and a grader can be compared against that agreement. Table G.5 sets the forward and bidirectional protocols against it.

The single forward disagreement occurs on an item whose grader verdict changed between runs. The published and re-run verdicts differ on it, so the forward protocol has no disagreement with the human consensus that is not run-to-run noise. The seven bidirectional disagreements are `basal_epithelia_rich` and `basal_epithelia_like` against *keratinocyte*, `Ganglion_cell_like` against *Neuron* and `stromal_cell_like` against *cancer-associated fibroblast*, where both adjudicators independently documented that one label represents a hypernym of the other, confirming that both names designate the same anatomical structure. The grader's instruction says to accept "a coarser/finer label". The forward call obeys it. The reverse call asks the same question with the terms swapped, which inverts the relation and fails it. A bidirectional-AND equivalence rule is therefore a granularity test as well as a correctness test, and on an open-vocabulary naming family that is a large effect.

A second block of 20 spatial items asked only whether the reference answers the question. Neither adjudicator answered yes. Their responses were 20 of 20 no, and 19 of 20 no with one unsure, with agreement 0.95. The keyword rule behind D2 reports 1 of 99 and two clinicians independently reach 0 of 20 on a sample of them.

H COMPUTE RESOURCES

The reported workload is evaluation, not a training study. The local workloads ran on L20 GPUs with 46 GB of memory each. Evaluating the fine-tuned and GRPO agent arms on the 120 questions took a few GPU-hours, and the probe on training questions took 99 agent runs. We trained no policy after the benchmark repair. The P0 checkpoint predates the repair, and its training data cannot be recovered. The frontier ceiling ran GPT-5.5 on OpenAI's hosted service through the Codex command-line client, at the client's default reasoning setting, which was not recorded. The 120 questions went out in five prompts of 9, 28, 28, 28 and 27 questions, each question with the executor's system prompt and its tool evidence, so answers within one prompt share a context.

I A COMPLETE EXECUTOR CONTEXT

Below is the verbatim string the agent's executor node receives for one niche-naming item, reconstructed by re-running the loop and recording its runtime context. The executor sees one line per completed node. Of the six lines preceding its answer, the fifth carries the evidence and the other five contribute a plan that is identical across questions, a region count with no masks attached, two constant embedding dimensionalities, and a literature retrieval that returned nothing. The critic that follows returns a constant score of one, so the loop accepts its first answer and never iterates.

```
Question: Identify the histological niche located near coordinates
  (4600, 5143) on slide NCBI267.
[planner] {'plan': ['segment tumor regions from H&E', 'compute
  niche-level expression']}
[segmenter] {'n_regions': 16}
[hist_encoder] {'dim': 1536}
[st_encoder] {'dim': 512}
[stats_tool] {'n_clusters': 5, 'clusters': [{'cluster_id':
  'leiden_0', 'centroid': [4793.7, 5443.6], 'top_genes': ['Fth1',
  'Mbp', 'Ctsd', 'Plp1', 'Apoe'], 'spatial_neighbors':
['leiden_3', 'leiden_1']}, {'cluster_id': 'leiden_1', 'centroid':
  [4201.3, 5428.1], 'top_genes': ['Snap25', 'Cplx1', 'Atp1b1',
  'Vamp1', 'Nat8l'], 'spatial_neighbors':
['leiden_0', 'leiden_3']}, {'cluster_id': 'leiden_2', 'centroid':
  [4104.7, 4868.4], 'top_genes': ['Map2', 'Snap25', 'Calm1', 'Zwint',
  'Meg3'], 'spatial_neighbors':
['leiden_3', 'leiden_1']}, {'cluster_id': 'leiden_3', 'centroid':
  [5073.1, 4881.1], 'top_genes': ['Plp1', 'Mbp', 'Fth1', 'Yam1',
  'S100b'], 'spatial_neighbors': ['leiden_0', 'leiden_2']},
  {'cluster_id': 'leiden_4', 'centroid': [4600.2, 5143.1],
  'top_genes':
['Apod', 'Vim', 'Dcn', 'Pmp22', 'Sparc'], 'spatial_neighbors':
  ['leiden_1', 'leiden_3']}], 'queried_location_cluster': 'leiden_4'}
[lit_rag] {'n_snippets': 0}
[critic] {'score': 1.0, 'commit': True}
```

The single-pass control of Section 5.2 is given the same question and the same evidence line under the same system prompt, and differs only by omitting the five non-informative lines.

J MODEL-FREE CHECKS ON BENCHMARKS BUILT BY OTHER GROUPS

To assess whether the diagnostics distinguish benchmarks, we ran three checks that need neither experts nor model calls, for answer leakage, for a uniquely longest correct option and for template concentration, on released files from other groups' agentic-science benchmarks (Table J.1). On these

Table J.1: **A length heuristic would score below chance in eight of nine subsets.** There the correct option is uniquely longest less often than chance. These model-free diagnostics use benchmarks we did not build, with LAB-Bench pooled over its eight subsets and the range across them. The last row is ours.

benchmark	n	answer leak (%)	correct option longest (chance, %)	largest template (%)
BixBench	205	2.4	8.8 (25.0)	1.0
LAB-Bench, 8 subsets pooled	1967	0.7	11.9 (23.8)	—
range over subsets	33–600	0.0–9.1	6.0–31.5	0.4–21.2
ours, spatial family	57	60	free text	40

benchmarks the checks rarely flag an item. Across BixBench and the eight LAB-Bench subsets, 1967 multiple-choice and 205 open items, answer leakage is 0.9% pooled (0.7% on the multiple-choice subsets, 2.4% on BixBench) and never above 9.1%, the correct option is uniquely longest on 11.9% of items against a 23.8% chance rate, and no subset is dominated by one template, whereas our spatial family leaks on 60% of items and draws 40% from one template. ProtocolQA is the only subset above its chance rate on the length cue, at 31.5% against 23.2%. One exclusion was decided after measurement. ScienceAgentBench first returned 87% leakage, but the detected string was the output filename its instruction must state, so we removed it from the leak column.

Answer leakage does not follow whether a reference is machine-made. DiscoveryBench provides human-curated and synthetic splits from the same authors, and the synthetic one leaks on 3.5% of 200 items against the human split's 9.6% of 239, a difference the exact test does not attribute to chance ($p = 0.013$).

K Replacing the grader on a benchmark we did not build

The analysis below supports the grader-replacement result in Section 5.4. The criteria, and what each possible outcome would mean, were fixed before any grading call.

BixBench selection. BixBench is the only external agentic-science benchmark we found that releases the agents' generated answers as well as the questions. Because the answers are fixed and public, the grader can be replaced without regenerating answers. Only the grader varies. We use its open-answer trajectories, 4,390 question items across the two image runs, each carrying a reference answer.

Each model is graded by itself. BixBench's released post-processing code selects the grader by matching the run name, so trajectories produced by GPT-4o are judged by GPT-4o and those produced by Claude 3.5 Sonnet by its dated snapshot of 2024-10-22. The paper describes a single judge. The judge on the GPT-4o side is requested by a floating alias with no dated snapshot, and their zero-shot grading path samples its judge at temperature one.

Reproducing their verdicts. Before any call, our re-implementation of their accuracy had to reproduce their own per-row verdicts from their released evaluation file. It does so exactly on all four open-answer runs, at 272 of 2,201, 261 of 2,219, 500 of 2,189, and 498 of 2,675.

Question-level pairing. Accuracy in Table K.1 is over the answer rows in the released evaluation file. The paired test is over questions. A question counts as correct for a model when a strict majority of its released replicas is graded correct. Ties count as incorrect. Dropping the 18 questions for which either model has no released row leaves 278. The capsule-level test is a signed-rank test on each capsule's difference in mean accuracy over its replicas, on the 50 capsules both models cover, and it agrees with the paired test in every row.

Verdict disagreement beneath a stable ordering. Table K.1 gives the result. The two aliases of one vendor, each reporting the same served model name, give different verdicts on 82 of 2,201 and

Table K.1: **BixBench's published ordering survives grader replacement, but its level changes.** Claude remains above GPT-4o under every grader tested, and its accuracy ranges over 4.5 points before rounding. The two DeepSeek aliases report the same served model but disagree on verdicts. Alias A is `deepseek-chat`, alias B is `deepseek-flash`, and the version-pinned open-weight grader is Qwen3.5-9B. All scores use released answers and count only non-empty rows, so the first row is 12.4% and 22.8%, whereas BixBench's all-attempt denominator of 2,960 gives 9.19% and 16.89% (last paragraphs of this appendix). b/c are discordant counts in the paired exact test over 278 questions.

grader	GPT-4o	Claude	gap (points)	b/c, p	parse rate
theirs, each model by itself	0.124	0.228	10.5	31/1, $< 10^{-6}$	—
alias A	0.139	0.219	8.1	24/2, 10^{-5}	1.000
alias B	0.110	0.184	7.3	20/0, 2×10^{-6}	0.988
open weights, version pinned	0.137	0.224	8.8	26/2, 3×10^{-6}	1.000

106 of 2,189 items, 3.7% and 4.8%. In 19 and 28 of these one alias returned a reply that could not be parsed, which the benchmark scores as an incorrect answer, and 63 and 78 differ where both replies parsed. The difference in answers credited as correct is smaller, 62 and 78, because disagreements in opposite directions cancel. Either alias regrading the same answers five days later changes 0.50% or 1.55% of its verdicts (next paragraph), so the two names differ from each other more than either differs from itself.

Graded twice. Each of the six grader and modality cells was graded a second time, five days after the first, with fresh caches and unchanged code, and the ordering holds in both passes under every grader. Verdicts changing between passes are 0.16% and 0.00% for the version-pinned open-weight grader, 0.50% and 0.59% for one hosted alias, and 1.55% and 1.37% for the other. For the pinned grader that figure isolates serving nondeterminism at temperature zero. For the hosted aliases, it combines sampling nondeterminism with any service change over those five days. These contributions cannot be separated, so we do not attribute it to either alone. The second pass was part of the protocol from the start and was run later than planned.

Grader replacement without images. Repeating the procedure on the two runs whose agents never saw figures, a further 4,894 question items, gives the same ordering result. The published gap of 6.9 points falls to 4.5, 4.3 and 4.5 under the three graders and reverses under none, every paired test below $p = 10^{-3}$ and the signed-rank agreeing throughout. This replicates the image result, so only that result is reported in the main text.

A grading protocol that excludes reasoning models. Their grading prompt ends by forbidding any output other than the verdict, and their parser accepts a single character. The reasoning configuration we tested cannot comply. Our open-weight grader, asked to reason first, returns its reasoning as ordinary prose with no delimiter to strip, and completes normally rather than running out of budget. Its parse rate is then zero, and because an unparsed reply is scored as the graded model being wrong, every item is marked incorrect. Without a parse-rate counter this is indistinguishable from a grader that scored the benchmark at zero. With reasoning disabled the same model returns a two-token verdict and parses on every item.

Unanswered questions leave the denominator. Their pipeline drops trajectories whose answer field is empty rather than scoring them incorrect. The rate differs across arms, from 6.9% to 22.0%, so the denominator depends on how often a system leaves its answer blank.

Their published accuracies, reproduced. Their released code divides the correct answers in each run by a fixed 2960, corresponding to 296 questions evaluated across 10 replicas, so a trajectory that failed or was never released and an answer left blank both count as wrong. Applied to their released verdicts, this calculation yields 9.19% and 16.89%, reproducing the 9% and 17% figures reported in their original paper. The higher nominal levels in our comparisons stem from evaluating all explicit candidate answers preserved in their released trajectory archives. Our protocol also fixed a third denominator, every question of every released trajectory with blank answers scored wrong, which the

Table L.1: **Source-supported model proposals do not establish a BixBench ranking change.** Accuracy is an equal-capsule mean over 29 capsules and 165 fixed answers per system. The final column is a within-capsule paired change in percentage points with a 95% capsule bootstrap interval. The DeepSeek request was served as `deepseek-flash`.

grader	answers of	released (%)	proposal (%)	paired change (points)
DeepSeek	Claude	6.8	6.0	-0.8 $[-3.8, 1.9]$
DeepSeek	GPT-4o	5.9	3.2	-2.7 $[-6.4, 0.0]$
Qwen3.5-9B	Claude	10.3	8.8	-1.5 $[-7.1, 4.1]$
Qwen3.5-9B	GPT-4o	6.2	4.5	-1.7 $[-3.6, 0.0]$

table does not use. It preserves the ordering under every grader, in both passes and both modalities, with every paired $p < 0.002$, while levels move by up to 4.1 points.

L External reference sensitivity on BixBench

The preceding analysis changes BixBench's grader on released agent trajectories. Here we change the reference while holding each answer and grader fixed. We use the original BixBench release, which has 296 questions in 53 source capsules (Mitchener et al., 2025). We include all capsules with an official source archive no larger than 20 MB, except two inspected in a feasibility pilot. This selection leaves 29 capsules and 165 questions. Archive size may select for simpler analyses, so these estimates do not represent all of BixBench. Each question has one released zero-shot open answer from GPT-4o (OpenAI, 2024) and one from Claude 3.5 Sonnet (Anthropic, 2024). These are not the agent trajectories used above. A second analysis later in this appendix scores those trajectories against the same references.

Before seeing the released references or either system's answers, GPT-6 Astra (OpenAI, 2026a), at Max effort, received each question, its capsule hypothesis and its executed source notebook. It proposed a reference with a notebook evidence pointer or abstained. It proposed 142 answers, marked eight questions ambiguous and found insufficient evidence for 15. After freezing those outputs, separate passes of the same model family checked each changed proposal against the source. Of the 142 proposals, 123 were supported as answers to the exact question, ten were unsupported and nine remained unresolved. Among the 123 supported proposals, 83 were substantively equivalent to the released reference, 39 differed in substance and one relation was uncertain. Unsupported and unresolved proposals retain the released reference. Source support is not human adjudication and does not establish that a different released answer is wrong. Human-reviewed BixBench revisions and an automated audit already exist (Phylo, 2026; Tu et al., 2026).

We score the fixed answers against the released reference, the source-supported proposal, a model-written rewrite constrained to preserve numbers and scientific qualifiers, and a different reference from a compatible question in the same capsule. The last control is eligible on 152 questions. We use BixBench's open-answer evaluation prompt with temperature zero and accept only a single binary verdict. Both graders parse all 1,294 requested verdicts. One is a DeepSeek endpoint requested as `deepseek-chat`, which reported `deepseek-flash` for every actual call. The other is a fixed local Qwen3.5-9B instance with its thinking mode disabled. The independent unit is the source capsule. Table L.1 gives equal-capsule means and 95% intervals from 10,000 paired capsule bootstrap samples. The released-answer accuracies here cannot be compared with Appendix K, which uses different answers and a different question population.

Claude 3.5 Sonnet remains ahead of GPT-4o under both references and both graders, with no within-capsule ordering reversal. The paired change in the Claude-minus-GPT-4o margin is $+1.9$ points (95% CI $[-0.6, 4.8]$) under DeepSeek and $+0.2$ points (95% CI $[-4.4, 5.4]$) under Qwen. Both intervals include zero. The question-weighted secondary margin changes are $+1.2$ and 0.0 points, respectively. On the 39 source-supported substantive alternatives across 21 capsules, the selected-subset margin changes are $+2.4$ points under each grader, with intervals $[-7.1, 15.1]$ and $[-7.9, 14.3]$. Only six distinct substantive questions change any verdict.

Under the DeepSeek grader, the wording control changes the score of the GPT-4o answers by -1.4 points (95% CI $[-3.6, 0.0]$) and leaves the Claude 3.5 Sonnet answers unchanged, and under the

Table L.2: **On agent answers the proposals narrow Claude's lead without reversing it.** Equal-capsule accuracy (%) under the released reference, and the paired change (Δ, points) with a pointwise 95% interval from 10,000 capsule bootstrap samples. The primary arm uses the 123 source-supported proposals and keeps the released reference elsewhere, the sensitivity arm uses all 142 proposals, and the physician arm replaces only the 24 references both physicians adopt. Margin is Claude minus GPT-4o, 4.8, 3.6, 8.0 and 8.2 points under the released reference in the four answer and grader cells. No-image rows cover 165 questions in 29 capsules and image rows 159 in 28. Qwen is the pinned Qwen3.5-9B.

answers	grader	Claude	Δ Claude	GPT-4o	Δ GPT-4o	Δ margin
primary arm, source-supported proposals						
no image	DeepSeek	21.8	−2.5 [−6.6, 1.0]	17.0	−0.9 [−2.9, 0.9]	−1.6 [−4.6, 1.2]
no image	Qwen	21.5	−1.4 [−4.7, 1.5]	17.9	−1.3 [−2.8, 0.0]	−0.1 [−2.4, 2.3]
image	DeepSeek	24.7	−2.8 [−7.2, 1.2]	16.8	−0.1 [−2.5, 1.9]	−2.6 [−5.9, 0.6]
image	Qwen	25.4	−2.0 [−5.6, 1.2]	17.2	0.0 [−1.5, 1.4]	−2.1 [−5.3, 0.9]
sensitivity arm, all proposals						
no image	DeepSeek	21.8	−3.2 [−7.2, 0.4]	17.0	−0.7 [−2.4, 1.0]	−2.5 [−5.9, 0.6]
no image	Qwen	21.5	−1.4 [−4.5, 1.5]	17.9	−0.7 [−2.2, 0.7]	−0.7 [−3.2, 1.9]
image	DeepSeek	24.7	−3.0 [−7.2, 0.8]	16.8	+0.4 [−2.1, 2.7]	−3.4 [−6.9, −0.1]
image	Qwen	25.4	−2.3 [−5.7, 0.9]	17.2	+1.3 [−0.6, 3.3]	−3.6 [−7.2, −0.2]
physician arm, the 24 proposals both physicians adopt						
no image	DeepSeek	21.8	−1.4 [−4.5, 1.0]	17.0	−0.2 [−1.2, 0.7]	−1.2 [−3.5, 0.7]
no image	Qwen	21.5	−1.0 [−3.6, 1.0]	17.9	−0.1 [−1.1, 0.8]	−1.0 [−2.7, 0.7]
image	DeepSeek	24.7	−1.4 [−4.4, 0.9]	16.8	−0.3 [−2.3, 1.0]	−1.1 [−2.8, 0.5]
image	Qwen	25.4	−1.1 [−3.2, 0.7]	17.2	+0.2 [−1.0, 1.2]	−1.3 [−2.9, 0.2]

Qwen grader it changes neither. The resulting margin change under DeepSeek is thus close to the full proposed-reference effect. The mismatched-reference control lowers the point estimate for each system and grader on its eligible questions. Its paired interval includes zero for three of the four cells. The proposal effect is small, grader-dependent and partly reproducible by wording alone. This experiment neither validates individual corrections nor shows that external benchmark rankings are generally unstable.

Power. Under the released reference both systems score at most 10.3% equal-capsule accuracy in any grader cell. An answer that is wrong under every plausible reference cannot change verdict. The zero-shot test therefore has little power on this population, and its null result says little about systems that answer more questions correctly. The 39 substantive alternatives are the questions on which a supported proposal happened to differ from the release. They were not sampled, and their margin change is descriptive.

Agent answers. A second protocol, fixed before any of its grading calls, keeps every reference arm, grader, prompt and parser above and changes only the answers. BixBench releases its agents' final open answers, several replicas per question, for four runs, the two systems each with and without image inputs. On the 165 questions, the no-image runs contain 1,494 and 1,235 answers and cover every question for both systems. The image runs contain 1,125 and 1,223 answers and cover 159 questions for both. A question's score is the mean verdict over a system's replicas, and blank answers score as wrong under every reference. The primary arm follows the rule of the zero-shot analysis above. A proposal replaces the released reference only where the source check supported it (123 questions), and a sensitivity arm uses the proposal on all 142 questions that have one. Table L.2 gives both. Accuracies lie between 16.8% and 25.4% under the released reference. In the primary arm, no system moves by more than 2.8 points. Claude's lead narrows by 0.1 to 2.6 points in every cell, every interval includes zero, and Claude stays ahead. The sensitivity arm narrows the lead by 0.7 to 3.6 points, and its two image-run intervals exclude zero. All 36 intervals in the table are pointwise and uncorrected, covering 12 margin changes and 24 accuracy changes. The wording-only rewrite moves no system by more than 0.8 points, and the mismatched reference lowers every system by 14.0 to 19.1 points, so the graders do read the reference on these answers.

Known-answer anchors. Each question has up to three anchor answers with a known relation to each reference, namely the paraphrase of the released reference, the proposal text and the mismatched reference. Both graders credit the paraphrase anchor on all 165 questions under the released reference and credit the mismatched anchor on 2.0% to 4.6% of its 152 questions under any arm. On the 54 questions whose proposal the source check calls substantively different (39 validated, 15 not), DeepSeek still credits the paraphrase anchor against the proposal on 24.1% and the proposal anchor against the released reference on 35.2%, and Qwen on 29.6% and 31.5%. Each grader therefore registers the change on 65% to 76% of the substantive items and treats the rest as equivalent, which limits how much a reference change can affect these anchor scores.

Physician review. Two postdoctoral physician researchers reviewed 43 items blind, under a protocol and an analysis script fixed before the sheets went out. They covered all 39 validated proposals that differ in substance from the released reference and four validated equivalent proposals as controls. For each item, they saw the capsule's hypothesis, question and relevant executed-notebook cells, including code and outputs. They chose which of two answers, labeled X and Y in an order randomized per item, was acceptable as the reference (X, Y, both, neither or cannot tell), with confidence from 1 to 5. They worked independently, without AI tools or web search, and were not told which was released. Both prefer the proposal on 24 of the 39 substantive items. On six more one prefers the proposal and the other cannot tell, on five one prefers it and the other accepts both answers, and on four one prefers the released reference and the other the proposal. Neither accepts the released reference on 30 items, and no item has both preferring it. The analysis counts cannot tell as accepting neither answer. Over the 43 reviewed items, controls included, κ is 0.341 $[-0.052, 0.732]$ for accepting the released reference and -0.126 $[-0.223, 0.000]$ for accepting the proposal, with capsule-bootstrap intervals. Both are low because nearly every choice falls on the proposal, and the item counts above describe the agreement better. Both accept both answers on all four controls. Against the 24 references both adopt, no agent score moves by more than 1.4 points and Claude's lead narrows by 1.0 to 1.3 points, every interval including zero (Table L.2), while the graders register the change on 91.7% and 95.8% of those items. Every proposal is longer than the released reference it would replace, usually because it carries its derivation, so a preference for a shown derivation cannot be separated from a preference for the value. The physicians are clinicians rather than bioinformaticians. The items were selected because a model proposal already differed from the release, so the 30 nonacceptances describe these items, not the frequency of incorrect references in BixBench. Of the 24 references they adopt, the two model reviewers below both adopt 12 and both keep the released reference on four. A third postdoctoral physician, whose sheet arrived before the cutoff, adjudicated the 15 contested items as the protocol allows. This does not change the two-reviewer figures above. The third physician preferred the proposal on nine, the released reference on four, neither on one and could not tell on one. A majority of the three then adopts the proposal on nine of the 15 and keeps the released reference on three, leaving three without a majority. As an exploratory sensitivity, replacing the 33 references adopted by both primary physicians or by a majority of three moves no agent score by more than 2.1 points and narrows Claude's lead by 1.0 to 1.5 points, every interval including zero.

Cross-family model review. Two models reviewed 75 items blind, all 59 whose proposal was not validated as equivalent and 16 controls, using the text prepared for human reviewers. GPT-6 Astra belongs to the family that proposed and checked the references, whereas Claude Opus 5.5 (Anthropic, 2026) is a second family. Each ran at Max effort without tools or web access. Each saw the question, the executed notebook's code and text outputs, and the candidates in random order, never which was released, and filled the fields of the human sheet under a rule fixed before any model answered. The models agree moderately, at $\kappa = 0.50$ $[0.26, 0.75]$ on accepting the released reference, 0.59 $[0.32, 0.88]$ on accepting the proposal and 0.75 $[0.62, 0.88]$ on how the two relate. Of the 39 validated substantive proposals both adopt 13, both keep the released reference on 9 and they split on 17, and counting each model separately, GPT-6 Astra prefers the proposal on 14 and Opus on 17. Of the 19 proposals the first family's own check did not validate, both adopt 2. Both keep the released reference on all 10 sampled equivalent proposals, and on 3 of the 6 items without a proposal neither accepts the released reference. Against the 15 references both adopt, no agent score moves by more than 0.6 points and the margin by at most 0.4, while the graders register the change on 86.7% and 93.3% of those items. These are model judgments, and the moderate agreement between two strong families shows that model review alone does not settle which reference is right.

Selection. Capsules were selected by archive size before any annotation, which may favor simpler analyses. Each proposal is checked against the notebook that produced the released answer, so an analysis choice made in the notebook but not stated in the question is reproduced rather than corrected. An independent audit of three BixBench capsules traced 73 of 80 incorrect agent answers to the answer key or the question's wording, including such unstated choices (Dai, 2026), and this analysis cannot detect that kind of defect. Twelve of our questions match the 50-question BixBench subset that Phylo reviewed with domain experts (Phylo, 2026), by capsule and identical released answer, because the subset's question text follows a later BixBench release that rewords most questions (the protocol and its one deviation are in the package). A crosswalk of 12 items describes these cases and is not an external validation of the proposals. The reviewers kept the answer on 11. On nine of them the proposal was validated as equivalent and on one no proposal was made. On the eleventh the reviewers rewrote the question to ask for the quantity the released answer reports, while the proposal rewrote the answer to fit the original question, so the two repairs address the same mismatch. For the one answer the reviewers corrected after specifying the analysis more fully, the proposal retained the released value. This is consistent with the notebook limitation described above.

M FULL LLM PROMPTS

The NicheLedger experiments used the following prompts. They were fixed before the main experiments. We did not tune them on the test set. The external reference experiment uses BixBench's released open-answer prompt, shown at the end of this section.

M.1 PLANNER PROMPT

```
System: You are the Planner for a spatial transcriptomics
agent. Given a biologist's question, output a JSON object
with a single field "plan": a list of 3-7 short atomic
steps. Do NOT execute them. Emit ONLY valid JSON, no other
text. Example: {"plan": ["segment tumor regions from
H&E", "compute niche-level expression"]}

User: {query}
```

M.2 EXECUTOR PROMPT

```
System: You are the Executor for a spatial transcriptomics
agent. You have access to the outputs of prior tools.
Produce (a) a short rationale (3-5 sentences), (b) an
<answer>...</answer> final answer. Do not call new tools.
Rely only on the prior observations.

User: Question: {query}\n[planner] plan=[...]\n[segmenter]
n_regions=7 label_hints=[...]\n[hist_encoder]
dim=1024\n[st_encoder] dim=512\n[stats_tool]
morans_i={niche: z-score}\n[lit_rag] snippets=[...]
```

The user message above is schematic. Appendix I gives the verbatim context one item received, with the values the tools actually returned.

M.3 GRADER PROMPT FOR SCORING AND BENCHMARK VALIDATION

```
System: You are a strict but fair grader for a
spatial-transcriptomics question-answering benchmark. You
are given a question, a reference answer, and a candidate
answer. Decide whether the candidate is SCIENTIFICALLY
EQUIVALENT to the reference for that question -- same
biological conclusion, even if the wording, granularity,
or ordering differs. Do not reward verbosity or penalise
correct synonyms. Reply with exactly one line:

VERDICT: CORRECT  or  VERDICT: INCORRECT

followed by a one-sentence justification.
```

```
User: {family line}\n\nQUESTION:\n{q}\n\nREFERENCE
ANSWER:\n{gold}\n\nCANDIDATE ANSWER:\n{pred}\n\nGrade the
candidate now.
```

The item's task family determines which of the three guidance lines is used.

```
Niche naming: This is a niche / cell-type identification.
Mark CORRECT if the candidate names the same or a
biologically equivalent cell type / niche as the reference,
even with different wording or a coarser/finer label.

Biomarker ranking: This is a biomarker gene-set question.
Mark CORRECT if the candidate lists a gene set that
substantially overlaps the reference (>=3 of the key genes)
OR is an accepted marker panel for the same cell type/niche.

Spatial relation: This is a spatial-pattern question. Mark
CORRECT only if the candidate identifies a niche with high
spatial autocorrelation AND gives a plausible biological
role AND cites a supporting gene set broadly consistent with
the reference. Exact JSON formatting is NOT required.
```

Every call uses temperature 0 and at most 256 output tokens. An empty candidate is scored incorrect without a call, and so is a call that fails three times. The forward call uses the reference and candidate order shown above. The reverse call swaps them, and the bidirectional rule credits an answer only when both return `CORRECT`. The parser takes the reply with any closed reasoning block removed, or the whole reply if that leaves nothing, and searches it without regard to case for the first occurrence of `VERDICT`, optionally followed by a colon or hyphen, and then `CORRECT` or `INCORRECT`. A reply with no such occurrence is scored as incorrect. The prompt asks for a `VERDICT` line rather than a bare boolean token. It does not prevent malformed replies, which the parser scores as incorrect and whose rate Section 5.1 reports for each grader name.

M.4 CoT-trace distillation prompt (SFT warm-up)

```
System: You are an expert spatial-transcriptomics
analyst. Think step-by-step, making each step an atomic
tool-call followed by its observation, and end with
<answer>...</answer>. Do not skip intermediate steps.

User: {query} (Slide: {slide_id}; H&E: {he_path}; AnnData:
{adata_path})
```

We set temperature 0.0 for the Planner, Executor, Judge, and validation roles, and temperature 0.7 for CoT-trace distillation.

M.5 BixBench open-answer grading prompt

The executable grading script preserves the released prompt string. The typesetting below normalizes whitespace.

```
Here is a question, the correct answer to the question, and
a proposed answer.
Question: {question}
Correct answer: {correct_answer}
Proposed answer: {proposed_answer}
You must respond with a binary score (0 or 1) for whether
the proposed answer is equivalent to the correct answer.
Nothing else is permitted.
```

Each call uses temperature zero and at most 16 output tokens. Only a reply consisting of `0` or `1`, apart from whitespace, is accepted. There were no empty candidate answers in the selected population.

Table N.1: **Naming accounts for most failures in the analyzed trajectories.** Observed failure modes on the analyzed T1 trajectories and on T3. The agent rarely fails to run a tool. It fails to turn a localized cluster into a biological name, and it fails on bordering-niche reasoning.

Failure mode	Rate
Output-format echo (raw `leiden_N` or `background` instead of a niche name)	64% of T1 answers
Segmentation misses the expression niche (region disjoint from gold)	12 of 33 T1
Naming error given the gold marker genes in isolation	∼76% (24% correct)
T3 spatial-relation reasoning failure (ceiling 2.5%, floor 0%)	near-total

Table O.1: **The seven arms of D4.** The fine-tuned policy's agent-loop and single-pass rows form the matched pair that isolates the scaffold contrast. All arms are graded under the identical strict protocol by the grader that produced the published figures. Table 2 gives the contrast under the two current graders. The last column excludes retrieval, whose prompt contains its own answer. †The GRPO arm predates the pipeline fix, has unrecoverable training data and is not interpreted.

	Success (%)				
arm	T1	T2	T3	overall	reasoning families
fine-tuned 9B in the agent loop	2.5	60.0	0.0	20.8	1.3
GRPO 9B in the agent loop†	5.0	55.0	0.0	20.0	2.5
fine-tuned 9B, single pass, thinking	17.5	45.0	0.0	20.8	8.8
base 9B, single pass, thinking	15.0	47.5	0.0	20.8	7.5
fine-tuned 9B, single pass, no thinking	12.5	35.0	0.0	15.8	6.3
base 9B, single pass, no thinking	10.0	27.5	0.0	12.5	5.0
frontier model, single pass	27.5	85.0	2.5	38.3	15.0

N FAILURE-MODE ANALYSIS

We analyzed the 33 niche-naming trajectories of the 99-trajectory probe of Appendix B, which ran the fine-tuned agent on training questions, together with the evaluation split of the spatial family. Because the probe uses training slides, its frequencies describe the agent on its training distribution rather than on held-out slides. Table N.1 lists the dominant failure modes, which lie in the output format and in the step from marker genes to a name rather than in the spatial statistics tools. Where the reference niche forms a ring or a scattered band, the segmenter is prompted at the cluster centroid, inside the hole, and returns a region that misses the niche.

The output-format echo can be removed, but performance elsewhere declines. Graded by the benchmark's grader, which credits one of the 33 original niche-naming answers that strict matching does not, an instruction forbidding the cluster identifier lifts niche-naming success on the probe from 3.0% to 9.1%, so part of the floor on that family is output format rather than missing knowledge. The same instruction lowers the retrieval family from 69.7% to 42.4%, leaves the spatial family at zero and lowers overall success from 24.2% to 17.2%, because the identifier was a shortcut the model used productively on retrieval. We keep the original prompt. That choice rests on counting the retrieval family, and a reader who sets that family aside would choose the modified prompt. The main-text replacement results hold each system's answers fixed and do not depend on this prompt choice. Six of the 99 queries returned no parsable answer under the modified prompt and count as incorrect, as unparsed replies do throughout. Excluding them gives 9.4%, 43.8% and 18.3%. Given the reference cluster's marker genes alone, the model names the niche correctly on 24% of queries. Clinicians reading marker lists reproduce the clustering reference at a similarly low rate (Appendix D), which is consistent with divergence from that reference but does not rule out a model deficiency. The two rates come from different graders and are not directly comparable.

O THE SEVEN ARMS OF D4

Table O.1 gives the success rates of the seven arms by task family.

P SEEDS, SPLITS AND SOFTWARE VERSIONS

The published agent-loop trajectories use seed 0 and the two repeats use seeds 1 and 2, respectively. The benchmark split uses seed 0, and the derangement control uses seed 20260908. The benchmark builder splits by slide rather than by question. Slides are grouped by platform and tissue. Within each group, they are ordered deterministically by a hash of the split seed and slide identifier, shuffled with a generator seeded by the split seed, and partitioned into 70% training and 10% validation, with the rest assigned to test. Groups with fewer than three slides go to training, and every question follows its slide. Rebuilding the split from the released question file at seed 0 reproduces the released training, validation and test files byte for byte, with no slide in two of them. Pinned software versions are listed below. The analyses use Scanpy (Wolf et al., 2018), and the agent serves its policies with vLLM (Kwon et al., 2023) and uses SAM 2 (Ravi et al., 2024), scGPT-spatial (Wang et al., 2025a) and the UNI pathology encoder (Chen et al., 2024b) as tools.

Package	Version	Package	Version
python	3.11.15	scanpy / anndata	1.11.5 / 0.12.16
torch	2.10.0	numpy / scipy / pandas	2.2.6 / 1.17.1 / 2.3.3
transformers	5.6.2	timm	0.9.16
peft	0.19.1	SAM2	commit `c2ec8e1`
trl	1.3.0	scGPT-spatial	commit `ed81440`
bitsandbytes	0.49.2	UNI checkpoint	UNI2-h, revision of 2025-07-01
vLLM	0.19.1	Qwen3.5-9B	revision snapshot `c202236`
langgraph	1.1.9		

These versions match the code package's requirements file, and random number generators (torch, numpy, Python) are seeded before each run. The package documents one command that recomputes every main-text number from archived predictions and grades and maps each number to its command or archived record. It cannot reproduce the hosted grading runs, whose grader is reached through an unpinned alias.