LITERATURERANK: BIAS-REDUCED LEADERBOARDS FROM SCIENTIFIC LITERATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce LiteratureRank, a literature-based approach for ranking methods in scientific domains that lack standardized leaderboards. The needed comparisons are often already published, but scores from different papers are not comparable, and most tables come from papers that propose one of the models they compare. We compare scores only within an evaluation cell, the scores that share a paper, a dataset and a metric, and expand each cell into pairwise wins and losses. To reduce author bias, self-report exclusion removes each paper's own proposed model from its tables during LLM-based extraction, so a model is ranked only from comparisons reported by other papers. A pair-weighted Bradley–Terry model turns these battles into scores with uncertainty intervals. We release 14 leaderboards and their comparison datasets, spanning general LLM and VLM capabilities and six scientific domains, which rank 1,082 models from 401k comparisons. A median of 74% of model pairs per board have non-overlapping 95% intervals, interval width shrinks as $n^{-0.38}$ in the number of comparisons, and widening one board's collection window by a year leaves the order of its existing models nearly unchanged (Spearman $\rho = 0.99$). The rankings can be updated as new papers appear with little additional computation.

1 INTRODUCTION

A leaderboard answers a question that the literature has, diffusely, answered many times already. When a pathology foundation model is published it is evaluated against several predecessors in the authors' own experiments; when a protein language model appears it is placed on ProteinGym (Notin et al., 2023) next to its ancestors; when a frontier LLM ships, a dozen independent evaluation papers put it beside its competitors within weeks. Each such table is a small, local, usually careful comparison. The field does not lack comparisons — it lacks a way to put them together.

Putting these comparisons together is of considerable value. Synthesizing evaluation results across papers is a metascience instrument — a way of studying scientific research itself (Ioannidis et al., 2015; Fortunato et al., 2018) — and funding agencies now support such work directly: the U.S. National Science Foundation, for instance, runs a dedicated Science of Science program aimed in part at evidence-based science policy (U.S. National Science Foundation, 2026). A cross-paper synthesis can inform such policy by showing which problems a field has largely settled and which remain open; it can help generate surveys of a field's results; and it can relate findings across papers without repeating the work behind them. Re-evaluating dozens of methods on every task a literature has used is tedious and expensive, and for methods without released code it is difficult to achieve at all. What is needed is a way to synthesize the numbers papers already report.

Two things stand in the way. The first is **incomparability**. A number in a paper is meaningful only relative to a dataset, a split, a metric, a unit and a protocol, and those differ from table to table; averaging accuracies across papers is not a summary but an artifact. The second is **advocacy**. Most evaluation tables appear in papers that propose one of the models in them. The proposed model's row is produced by its own authors, under their own tuning budget, against baselines they chose; the documented consequence is that proposed models win (Marie et al., 2021; Dehghani et al., 2021). A naive aggregation of published numbers ranks models by how recently and how enthusiastically they were published.

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.

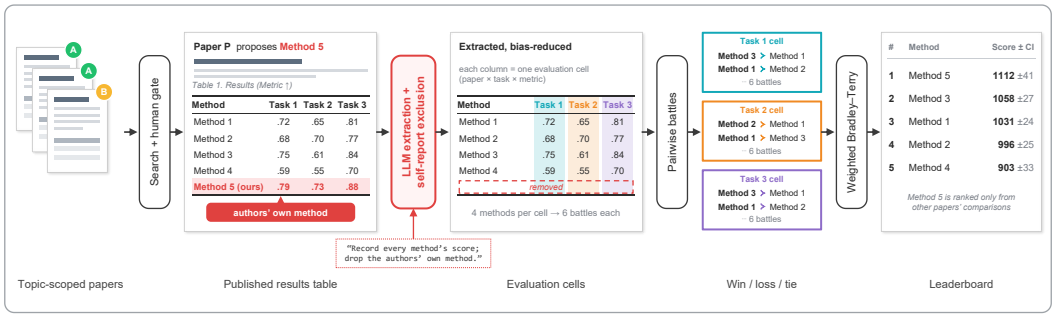


Figure 1: The LiteratureRank pipeline, shown on a schematic paper that proposes Method 5. (1) A gated review workflow produces a topic-scoped paper set in any domain (§3.1). (2) An LLM extracts every reported score and, as the bias-mitigation step, removes the authors' own method from that paper's records (red; §3.2). (3) Each column of the remaining table — scores sharing paper, task and metric — is an evaluation cell, and (4) a cell of k methods expands into $\binom{k}{2}$ win/loss battles (§3.4). (5) A pair-weighted Bradley–Terry fit over the battles of all papers produces the leaderboard (§4); Method 5 is still ranked, but only from comparisons reported by other papers. Names and numbers are illustrative.

Chatbot Arena (Chiang et al., 2024) solved a structurally similar problem for chat models by giving up on absolute scores: it collects *pairwise* preferences, which need no common scale, and fits a Bradley–Terry model (Bradley & Terry, 1952) to recover one latent strength per model. The published record admits the same reading. Within a single evaluation table two numbers *are* comparable — same data, same metric, same protocol — even though neither is comparable to a number in another paper. A table with k models is therefore not one measurement but $\binom{k}{2}$ pairwise outcomes, and the union of all tables in a field is a comparison graph of the kind Bradley–Terry was designed for. This lets us sidestep the incomparability between papers. We call a set of scores that share a paper, a dataset and a metric an **evaluation cell**, and compare numbers only inside the cell that measured them together: no score is ever converted, rescaled or averaged across papers.

This reframing also gives a handle on advocacy. If a model paper's bias lives mainly in the row describing its own contribution, deleting that row converts an interested source into a largely disinterested one: the authors of a pathology model have every reason to make their own numbers strong, but little reason to distort the relative ordering of the ten baselines they cite. We make this the organizing principle of extraction and call it **self-report exclusion**.

Evaluation cells make each comparison locally sound, but they do not make papers interchangeable. A ranking still chains together comparisons made on different data, under different tuning budgets and protocols, and so assumes that a method's relative strength holds across settings. We scope LiteratureRank accordingly. It is not a substitute for a controlled benchmark, and a board does not certify that one method beats another. Its purpose is to surface field-level trends — which methods the literature consistently places above which, and where that evidence is thin — and such metatrends remain useful even though they are not validated in the way the underlying scientific findings are.

Contributions. **(i)** An end-to-end auditable literature-to-leaderboard pipeline (Figure 1): every battle traces back to a named evaluation location in a named PDF, and every intermediate artifact is released. **(ii)** A *reusable discovery skill*: Phase 1 is an agent skill driven by five user inputs rather than re-implemented per domain, and it stops itself at a human review gate (§3.1). **(iii)** *Self-report exclusion* as an extraction-time debiasing device: a paper-type taxonomy plus a deterministic variant-matching filter that holds even when the extracting model disobeys the prompt (§3.2). **(iv)** The *evaluation cell* as the unit of comparability, reducing heterogeneous tables to a homogeneous battle stream with no cross-paper unit conversion (§3.4). **(v)** A Bradley–Terry fit adapted to sparse, directionally structured literature graphs: an evidence k-core filter and Ford's (1957) strong-connectivity requirement, alternated to a fixed point (§4.2). **(vi)** 14 released leaderboards and their datasets: 1,082 ranked models over 401k retained comparisons (§5). **(vii)** Two measurements of what a board is worth: interval width scales as $n^{-0.38}$ rather than $n^{-1/2}$, and a widened collection window extends a board without reordering it ($\rho = 0.99$; §6).

2 Related Work

Pairwise-comparison leaderboards. The Bradley–Terry model (Bradley & Terry, 1952) gives each competitor a single hidden strength: the stronger of two competitors is more likely to win, and more so the larger the gap. Ford (1957) showed when a set of outcomes pins these strengths down uniquely — roughly, when no group of competitors has won, or lost, every comparison against the rest. Later work extended the model to ties (Davidson, 1970) and made it fast to fit (Hunter, 2004; Newman, 2023); the Elo ratings used in chess (Elo, 1978) update the same kind of strength one game at a time. Chatbot Arena (Chiang et al., 2024; Zheng et al., 2023) applied the model to crowdsourced human votes between chat models, and we adopt its way of reporting results: a score per model, an uncertainty interval around it, and the range of ranks each model could plausibly hold. Our departure is the source of the comparisons: published evaluation tables, not anonymous users.

Static and holistic benchmarks. MMLU (Hendrycks et al., 2021), BIG-bench (Srivastava et al., 2023), SWE-bench (Jimenez et al., 2024) and HELM (Liang et al., 2023) standardize evaluation by fixing the data and protocol — the complementary strategy: they achieve comparability by controlling the measurement, we by only ever comparing numbers already measured together. Their failure modes are saturation, contamination and benchmark lottery (Dehghani et al., 2021; Raji et al., 2021; Recht et al., 2019; Kapoor & Narayanan, 2023); ours is selection in what the literature chose to compare (§7).

Extracting results from papers. Hou et al. (2019) framed leaderboard construction as extraction of (task, dataset, metric, score) tuples and AxCell (Kardas et al., 2020) built a table-parsing system for the same target. LLM-based structured extraction from scientific text is now effective in specialized domains (Dagdelen et al., 2024). These systems aim to *reproduce* each paper's reported numbers; we aim to *aggregate* them, while recording the provenance of every row — who reported which result.

Meta-analysis and meta-evaluation. Aggregating published estimates is the business of meta-analysis (Glass, 1976), with established vocabulary for heterogeneity (Higgins & Thompson, 2002), publication bias (Egger et al., 1997) and review protocol (Page et al., 2021). Marie et al. (2021) meta-evaluated 769 machine-translation papers and documented systematic distortions in reported comparisons; Bouthillier et al. (2021) quantified the variance benchmark numbers carry. LiteratureRank is a meta-analysis whose effect measure is a pairwise win rather than a standardized mean difference, buying immunity to cross-study scale differences at the cost of effect magnitude.

3 From Papers to Comparisons

3.1 Topic-scoped literature collection

The workflow runs in two phases separated by a hard human gate. Phase 1 is released as an agent skill: an agent invokes it, it drives MCP search servers over the public metadata sources, and it terminates itself at the gate. Nothing about it is specific to a domain.

Five inputs, then a frozen configuration. The skill takes exactly five user inputs — a *topic* in the user's own words; one or more *object groups* (e.g. "protein language models", "protein benchmarks"); the *input* modality the objects require; the *output* they produce; and a *time* window given as a cutoff or an interval, which becomes the board's collection window in Table 1. There are no built-in domain vocabularies, and a missing input is asked for rather than guessed. Inputs two through four are composed into one sentence, the *target capability*, which becomes the literal criterion for the second strict condition below. Initialization freezes the configuration, the scope snapshot and a search-plan skeleton into a run directory with content hashes, and changing any input, term list or rule requires a new run.

Plan, then search. The search plan — concepts, synonyms and exclusion terms, query combinations, pagination, deduplication keys and stopping conditions — is written before any query is issued, so that an auditor can reproduce the run. Two tracks, *models* and *benchmarks*, are searched,

deduplicated and graded independently, then joined. arXiv, Crossref and OpenAlex are the discovery sources; each query is paginated to exhaustion or a protective limit, and raw responses are stored with their parameters and actual stopping reason.

Three strict conditions, then a gate. A candidate must be (1) a *new public object* first released, versioned or proposed after the cutoff; (2) a verbatim match to the *target capability* sentence; and (3) identified by a *first-hand primary record*, for which an arXiv or venue record suffices while aggregator listings do not. Each condition requires quoted source evidence with its location; a keyword hit is never a substitute, and thin recall is remedied by revisiting query terms, never by relaxing the capability condition. Phase 1 then stops at the terminal state `awaiting_user_review` without downloading full text; a validator checks that accepted rows carry all three pieces of evidence, and approval is logged against the run, report and configuration hashes. **Phase 2** downloads the full text of approved papers only and passes each to extraction (§3.2); Appendix A gives the details.

3.2 Extraction with self-report exclusion

Each approved PDF is sent to a language model under a single JSON-only prompt. We use two models in total to assist extraction — Claude Opus 5 at low reasoning effort and GPT-5.5 at medium — and the raw response is cached per PDF, so re-normalization never re-queries the API and every downstream row traces to a stored response. Appendix B reproduces the prompt.

Paper typing. The model first assigns `paper_type` $\in$ {`model_paper`, `benchmark`, `model_and_benchmark`, `other`}: a paper that proposes, trains or releases a focal system; one that primarily introduces or evaluates a benchmark, dataset, evaluation suite, protocol or leaderboard; one that does both; and one that does neither, which returns no records.

The exclusion rule. For `model_paper` and `model_and_benchmark`, the prompt requires the exact focal model name to be read off the title, abstract, method or tables — never inferred from the filename — and returned in a top-level field. Every evaluation of that model is then dropped: case and spelling variants, dated or size variants, configuration variants, checkpoints and fine-tuned versions of the same release, and the labels `Ours`, `Our model` and `This work` wherever they head a row. All *other* deep-learning models in eligible blocks are kept, ablations and component-removed variants aside. For `benchmark`, the focal-model field is empty and every reported model is recorded, baselines included, with an explicit instruction to keep non-headline rows.

Defense in depth. Prompt compliance is not assumed. After parsing, a deterministic filter re-derives the same decision: names are reduced to alphanumeric keys and a candidate is focal when its key equals, prefixes, or is prefixed by the focal key, or when its label is a generic self-reference. Rows failing this test never reach the normalized table but stay in the cached response, so the audit trail records what the extractor proposed and the release what survived.

Scope discipline. The prompt is strict on input modality and open on output: for pathology the evaluated system must take a histopathology image, while any evaluated task counts. Raw values are kept verbatim — no unit conversion, winner inference, ranking or estimation from plots unless an exact value is printed — and a valid record needs an identifiable model, a benchmark/task context, a metric and a reported numeric score.

3.3 Normalization

Values are parsed with one grammar accepting an optional percent sign and an optional $\pm$ dispersion term, retaining mean and dispersion separately. **Metric direction** is resolved in a fixed precedence — the extractor's own flag, an override table for names that read the wrong way (`fnr`, `params (m)`), context-sensitive regular expressions over the benchmark, task and dataset fields, then keyword lists — with the context rules ahead of the keyword lists so that an adversarial quantity such as "attack success rate" is not read as a desirable success rate (Appendix C). Metrics with no direction at all — a length, a fairness spread — are placed on an unordered list and their rows removed before any battle is formed, rather than ranked under a guessed sign.

Model identity. Papers name the same checkpoint many ways (`UNI2` / `UNI-2` / `UNI2-h`; Chen et al., 2024), attach citation tags (`TransMIL [53]`) and usage notes (`CONCH (frozen)`; Lu et al., 2024). These collapse onto one identity through a catalog of canonical names, aliases and ordered regular expressions, ordered so that a checkpoint is never absorbed by a rule written for its own prefix. Two domain decisions matter. A slide-level row that names an encoder and its aggregator (`UNI + ABMIL`) is attributed to the encoder, the object these boards rank; the aggregator is a small head that tables typically hold fixed. And labels naming no single model are dropped: `CNN` is a three-layer network in one paper and a fine-tuned ResNet-18 in the next, and `Baseline` and author handles fail the same way. Appendix C lists the rules.

3.4 Evaluation cells and battles

The unit of comparability is the *evaluation cell*: the tuple (source collection, source PDF, paper title, benchmark, task, dataset, split, metric, unit, evidence location). Two scores in the same cell were produced by the same authors, on the same data, under the same metric and protocol, and reported in the same place. We assume nothing about two scores in different cells. Within a cell, duplicate model rows are collapsed and the k remaining models expand into all $\binom{k}{2}$ battles, with outcome $y_{ij} = 1$ if model i is better on this metric, 0 if j is, and $\frac{1}{2}$ if the reported values are exactly equal. Cells with fewer than two models contribute nothing.

No magnitude crosses the cell boundary: a 0.3-point win and a 30-point win are the same battle. The loss is deliberate — it is what makes a histology AUROC table and a captioning BLEU table summable at all, and why the resulting scale is a strength ordering, not an accuracy estimate.

4 Ranking

4.1 Weighted Bradley–Terry

Section 3.4 reduces the corpus to n battles $(i_r, j_r, y_r)_{r=1}^{n}$ over M models, where $y_r \in \{0, \frac{1}{2}, 1\}$ is the outcome for i_r. Write $x_r = e_{i_r} - e_{j_r} \in \mathbb{R}^M$ for the design vector of row r and $\sigma(z) = (1+e^{-z})^{-1}$. Bradley–Terry gives each model a latent strength $\beta \in \mathbb{R}^M$ and models a battle as

$$\Pr(y_r = 1 \mid i_r, j_r) \;=\; \sigma(\beta_{i_r} - \beta_{j_r}) \;=\; \sigma\big(\beta^\top x_r\big)\,. \tag{1}$$

Since $\beta^\top x_r$ is invariant to $\beta \mapsto \beta + c\mathbf{1}$, the strengths are identified only up to a shift; we fit under the reference coding $\beta_1 = 0$ and report the centered representative $\sum_i \beta_i = 0$. With $\ell(y,p) = -[\,y \log p + (1-y)\log(1-p)\,]$ the binary cross-entropy, evaluated at $y = \frac{1}{2}$ for a tie so that each side is credited half a win, the board is the weighted M-estimator

$$\hat\beta \;=\; \operatorname*{arg\,min}_{\beta:\,\beta_1=0} \;\sum_{r=1}^{n} w_r\, \ell\big(y_r,\, \sigma(\beta^\top x_r)\big)\,, \qquad w_r \;=\; \frac{n}{m_{p(r)} \sum_{s=1}^{n} 1/m_{p(s)}}, \tag{2}$$

where $p(r) = \{i_r, j_r\}$ is the unordered pair of row r and m_p its multiplicity. Ties are not dropped and no tie parameter is estimated, so this is not a Davidson model (Davidson, 1970).

Why the weights. An arena samples pairs from a known P and can invert it, weighting a vote by $1/P(A_t)$ to target the score under a uniform distribution over pairs (Chiang et al., 2024). The literature supplies no P: one popular cohort or one reused table template can report a pairing hundreds of times, which an unweighted fit reads as hundreds of independent observations. The empirical pair frequency m_p/n takes its place, and the w_r of (2) are the corresponding inverse-probability weights, rescaled so that $\sum_r w_r = n$: each unordered pair contributes the same total weight however many tables reported it, so no single source can dominate the board — the practical counterpart of the *resilience* Fageot et al. (2024) formalize for this family. Weights are recomputed after every filtering step of §4.2.

Optimization. The objective in (2) is convex, and we minimize it by damped Newton steps under a pseudo-inverse Hessian with backtracking, accepting a run only if it converged *and* the weighted score equations are solved to $\|g(\hat\beta)\|_\infty < 10^{-5}$ (Appendix D gives g, H and the step). Strengths go on the Arena scale $s_i = 1000 + (400/\ln 10)\,\hat\beta_i$, so the board mean is 1000, a 100-point gap is 64:36 odds and a 400-point gap is 10:1.

4.2 IDENTIFIABILITY ON LITERATURE GRAPHS

Ford's condition (Ford, 1957) states that the Bradley–Terry MLE exists and is unique exactly when the win graph is strongly connected: for every split of the models into two non-empty sets, someone on each side must have beaten (or tied) someone on the other. Chat-arena graphs satisfy this easily, because pairs are sampled at random; literature graphs do not. Whole families are compared only in one direction against the rest — classical CNN baselines that every modern paper cites and every modern paper beats form a separated block — and the fit drives that block's parameters to the boundary while the gradient quietly goes to zero.

We therefore restrict the fitting set before estimation. Two restrictions interact, and each can break the other, so they are alternated to a fixed point (Algorithm 1, appendix). The **evidence k-core** requires a model to appear in at least $\tau = 50$ comparisons, applied twice for different reasons: once on the full canonicalized corpus *before* any allowlist, which asks how much evidence the literature contains about a model — a property of the model, not of who else is on the board — and then iteratively *inside* the fitting set, because a model can clear the corpus threshold on battles against unlisted models and still enter the fit on a handful of rows, landing at an extreme score. **Strong connectivity** keeps the largest strongly connected component of the directed win graph, computed with an iterative implementation of Tarjan's (1972) linear-time algorithm; ties contribute an edge in both directions. This is strictly stronger than the undirected connectivity a naive implementation checks, and it is the restriction literature graphs actually need.

Within each round the count filter runs first, so "is there enough evidence about this model" is answered on the comparisons that exist rather than on a set already pruned for a structural reason. The count filter's fixed point is unique (it is a k-core), so only the choice of component is order-sensitive, and the loop terminates because the model set strictly decreases whenever it changes. Appendix D reports what the restriction removes on each board.

Two regimes. The allowlist A is used on nine of the fourteen boards and not on five, and the split is not the LLM/science split. For LLM and VLM capabilities we maintain a curated list of released systems whose identity, vendor and version we can resolve, so those boards rank a fixed, auditable population that no mis-transcribed or hallucinated name can join. Pathology is fitted the same way, against a roster of 64 released encoders whose weights, paper and version we can point at: the device needs only a population whose membership is decidable in advance. The other five scientific boards have none — there the population is what we are trying to discover, and fixing a list would predetermine the answer — so they admit every model the literature names and the evidence k-core does the list's work, at the cost of a longer tail and heavier canonicalization (§3.2). Table 2 records the regime of each board.

Why restrict rather than regularize. Ford's condition constrains the *unpenalized* MLE, not the Bradley–Terry model, and extensions relax it by perturbing the win matrix, penalizing the likelihood, or adding structure (Conner & Grant, 2000; Yan, 2016; Fageot et al., 2024; Firth, 1993; Kosmidis & Firth, 2021; Turner & Firth, 2012; Cattelan, 2012). We use none of them: each supplies the missing ordering from a penalty rather than the record, so on a sparse board much of the ranking would be an artifact of a constant we chose. Restriction instead pays for identifiability in coverage, a loss we report exactly (Appendix D).

4.3 UNCERTAINTY AND RANK SPREAD

Battles are not independent: one table contributes $\binom{k}{2}$ correlated rows, and one cohort may appear in dozens of papers, so the model-based H^{-1} would be optimistic. Because (2) is an M-estimator, its sandwich variance remains valid when (1) is misspecified (Huber, 1967; White, 1980), and we report its HC1 form (MacKinnon & White, 1985)

$$\widehat{\mathrm{Var}}(\hat{\beta}) \;=\; \frac{n}{n-p}\, H(\hat{\beta})^{-1}\Big(\textstyle\sum_r u_r u_r^{\top}\Big) H(\hat{\beta})^{-1}, \qquad u_r = w_r\,(\sigma_r - y_r)\,x_r, \tag{3}$$

with $p = M - 1$ free parameters, propagated through the centering transform onto the Arena scale. Following Chatbot Arena we report a *rank spread* rather than a point rank: with $[L_i, U_i]$ the 95% interval of model i,

$$R_i^{\mathrm{best}} = 1 + \left|\{\, i' : L_{i'} > U_i \,\}\right|, \qquad R_i^{\mathrm{worst}} = M - \left|\{\, i' : U_{i'} < L_i \,\}\right|, \tag{4}$$

Table 1: The 14 released leaderboards. *Window* is the publication-date range of the papers the board draws on, a Phase 1 input (§3.1); it bounds when a comparison was published, not when the models in it were released. *Input* is the candidate battles produced from extracted evaluation cells, *fitted* the number surviving the restrictions of §4.2; *density* is observed pairs as a fraction of $\binom{M}{2}$ and *votes* the median fitted comparisons per model. The two vision rows are nested windows over one track (§6.3), so the totals double-count it.

Leaderboard	Window	Models	Input battles	Fitted	Pairs	Density	Votes
General LLM / VLM capabilities							
Reasoning & Knowledge	2026	47	200,904	19,149	439	41%	522
Math	2026	88	61,938	11,543	3,828	100%	202
Coding	2026	82	54,551	27,666	3,048	92%	292
Agentic	2026	84	235,721	91,579	2,938	84%	801
Instruction Following	2026	81	53,021	16,629	2,831	87%	406
Multilingual	2026	41	138,955	5,460	498	61%	206
Vision (2025–26)	2025–2026	73	643,591	78,491	1,504	57%	1,340
Vision (2026)	2026	48	431,051	38,689	835	74%	883
Scientific domains							
Pathology foundation	2019–2026	63	181,839	51,269	650	33%	398
Protein language	2026	158	36,436	24,174	1,607	13%	213
Remote sensing	2026	117	24,402	11,758	848	12%	94
Agriculture	2026	95	17,050	11,750	740	17%	204
Chemistry / drug disc.	2025–2026	53	9,782	3,605	244	18%	112
Spatial transcriptomics	2019–2026	52	9,847	9,148	341	26%	220
Total	–	1,082	2,099,088	400,910	20,351	–	–

so a model is ordered only against the models it separates from.

5 Corpus and Leaderboards

Table 1 summarizes the 14 released boards. They divide into general LLM/VLM capability boards, where the population is frontier models and the literature is thick with independent evaluation papers, and scientific domain boards, where the population is domain models — pathology foundation models (Chen et al., 2024; Lu et al., 2024; Xu et al., 2024), protein language models (Lin et al., 2023), and the populations around MoleculeNet (Wu et al., 2018), EuroSAT (Helber et al., 2019) and PlantVillage (Hughes & Salathé, 2015) — and most evidence comes from model papers stripped of their focal rows. Scale is the dominant structural fact: the dense boards rank 41–88 models on near-complete graphs (Math observes every one of its $\binom{88}{2}$ pairings, Coding 92%), while the open-population scientific boards rank many more on far sparser ones — 158 models at 13% density on protein language, 117 at 12% on remote sensing. The identifiability machinery of §4.2 is close to inert on the first and load-bearing on the second. Figure 2 shows the top of two boards; Appendix G gives the top ten of all fourteen, and Appendix E reproduces one board in full.

Density reports the population, not the field. The gap follows from the allowlist decision of §4.2, not from how much each field cross-compares: on a listed board every paper draws baselines from the same few dozen systems, so the graph comes out dense, while an open board's population grows with the corpus and the same evidence covers a larger pair space (Appendix D).

Pathology is the controlled version of that claim, because we have run it both ways on one corpus. Fitted open, the same 181,839 input battles put 484 models on a graph observing 3.7% of pairs. Fitted against the field's roster of released encoders — the same device the LLM boards use — it is 63 models, 51,269 comparisons, 33% density, a median 398 votes per model, and 73% of pairs separated at 95%, inside the general-capability range rather than the open scientific one. Only the admitted population changed between the two fits; the released board is the listed one.

That sparsity is not fixed. Every board ships its fitted comparison set and model catalog, so imposing a different population is the edit we made for pathology: filter to a cohort and refit, and the loop of §4.2 reports its own coverage loss (on pathology, 455 models below the floor and 92,764 rows

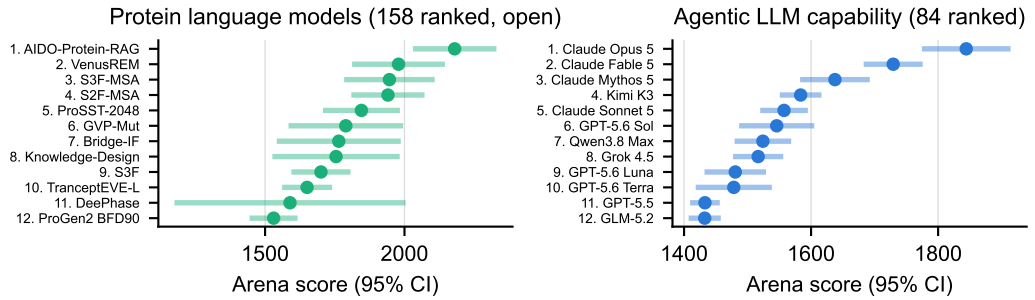


Figure 2: Top-12 entries of one open-population scientific board and one listed general-capability board, with 95% HC1 intervals. The dense board separates its leaders; the sparse protein-language board does not.

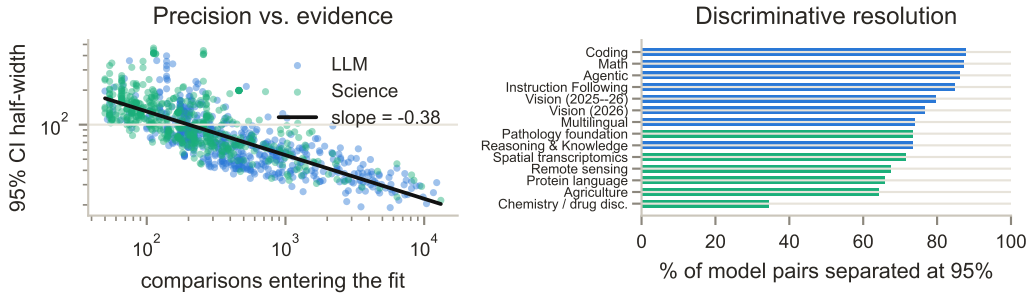


Figure 3: Left: interval half-width against fitted comparison count, all 1,082 ranked models; the fitted slope of -0.38 is shallower than the -0.5 of independent observations. **Right:** fraction of model pairs per board whose 95% intervals do not overlap.

naming an unlisted model). The 14 boards are one defensible population per topic, not the only possible fit.

The windows are mostly narrow: ten boards draw only on 2026 papers, two on 2025–2026, and two — pathology and spatial transcriptomics — reach back to 2019. A window bounds when a comparison was *reported*, not when its models were *released*: the 2026-only agriculture board still ranks AlexNet and VGG-16 alongside Gemini-3-Flash. Spatial transcriptomics has no benchmark papers in its approved set, so self-report exclusion carries the whole debiasing burden there.

6 Analysis

6.1 What is a literature battle worth?

For independent binary comparisons, the standard error of a Bradley–Terry strength falls as $n^{-1/2}$. Literature battles are not independent, and the departure is measurable: regressing log interval half-width on log fitted comparison count over all 1,082 ranked models gives a slope of -0.38 (Figure 3, left) rather than -0.5.

Read as an effective sample size, n literature comparisons carry roughly the information of $n^{0.76}$ independent ones; a model with 1,000 fitted battles is worth about 190 independent votes. This is the price of the cell structure: a table's $\binom{k}{2}$ battles share one set of authors, one data preparation and one tuning budget, and pair weighting removes only the most blatant part of it.

6.2 How much does a board actually resolve?

A leaderboard that cannot separate its entries is a ranked list, not a measurement. Across the 14 boards a median of 73.7% of model pairs are separated at 95% (Figure 3, right), from 34.6% (chemistry, the smallest corpus) to 87.8% (coding). The split runs with the population regime rather than

with the subject: the nine listed boards, pathology among them, cluster between 73% and 88% with a median interval width of 106 Arena points, while the five open boards run from 35% to 72% at a median width of 224. The median interval across all boards is 146 points, about $1.5\times$ the 100-point gap that is $64{:}36$ odds (§4.1), so resolution is a pairwise property, read off interval overlap. By (4) a unique rank needs separation from *every* other model, which on boards this long almost never holds: the informative statement is the spread — the 158-model protein-language board lacks the evidence to order its middle.

6.3 What happens when the window widens

The vision track was run over two nested, otherwise identical collection windows — 2026 papers alone, and 2025 onward — which shows what happens to a board as the literature accumulates. Widening the window adds models without disturbing those already ranked. The boards share 48 models, and we compare them on that shared set alone: each board's fitted scores are re-ranked $1 \ldots 48$ within it, so the 25 added models cannot inflate the agreement by pushing everyone down together, and the two rank vectors are then correlated. They agree at Spearman $\rho = 0.991$ and Kendall $\tau = 0.936$; the median model moves one position, no model moves more than five, and among the 48 shared models the top ten is the same set on both boards (Figure 5, appendix). The added models are predominantly older and land where that implies: median rank 53 of 73, with 80% in the bottom half. Precision improves at the same time — on the shared models the median 95% interval narrows from 104 to 90 Arena points, a 13% tightening bought purely by a year of further papers about the same systems. The cost of a narrow window is reach, not error: the 2026-only board cannot speak about the 25 models it stopped re-benchmarking.

7 Limitations

This is not human preference. The machinery is Arena's; the data is not. No anonymous user chose anything, and we ran no ground-truth evaluation. Style control, direct-battle indicators and organization covariates cannot be estimated, because the underlying fields do not exist.

Selection in what gets compared. A board inherits two selections. Whole evaluation dimensions can be missing — if a community never reports calibration or never tests under distribution shift, nothing on that axis can be aggregated, and nothing on the board says which questions went unasked. And the choice of opponent is not random: authors run baselines that are conventional, cheap to reproduce or flattering, so a widely cited model becomes rankable while a strong but unfashionable one may never be compared; the evidence k-core makes this explicit rather than fixing it.

Residual advocacy. Self-report exclusion removes the focal model's own rows, not an author's choice of *which* baselines to run or how hard to tune them. Under-tuned baselines depress the remaining rows together, and a pairwise model cancels this only to the extent that the under-tuning is uniform across them.

Magnitude is discarded. Win/loss/tie is what makes heterogeneous tables summable, but a board cannot say how much better a model is in units a practitioner cares about; and only exact equality registers a tie, so an indistinguishable 0.1-point difference reads as a decisive win.

Extraction, identity and temporal mixing. Paper typing, focal-model identification, metric direction and name canonicalization are automated and can all fail; direction heuristics are the most exposed, since a wrongly resolved metric inverts every battle in its cell, so we release the direction source for every row. And a board mixes model generations even when its window is narrow, so static strength describes the archive, not a head-to-head today.

8 Conclusion

The published benchmark record is a large, partly adversarial, highly structured source of pairwise information about model quality. Read as a comparison graph rather than a pile of numbers, it yields leaderboards auditable back to the page they came from: across 14 boards the scales separate three quarters of model pairs, survive a widened window at $\rho = 0.99$, and carry an $n^{-0.38}$ law for what a literature comparison is worth.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: implementing methods, cleaning and reformatting the dataset, and supporting qualitative data analysis. Large language models are a component of the method itself. Claude Opus 5 and GPT-5.5 type each paper and extract its reported results into structured records (§3.2, with the prompt reproduced in Appendix B), and an LLM agent executes the Phase 1 skill that searches for candidate papers and screens them against the inclusion criteria (§3.1). We also used generative AI tools to help implement the analysis code. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing ingredients for proving mathematical claims, writing proofs, proposing or refining hypotheses, designing or giving feedback on the research methodology or experiments, translation, or interpreting results. Additionally, we used generative AI tools for the following tasks with recommended disclosure: searching for and identifying relevant literature (the Phase 1 skill above), creating and editing software code, creating figures, and drafting and editing the paper text.

We have reviewed all AI-assisted work. LLM outputs used by the method are not trusted directly: extracted rows pass a deterministic self-report filter and schema validation, Phase 1 stops at a mandatory human review gate, and every extracted row traces to a cached raw response that is released. Text was read and revised by the authors, code was checked against the released artifacts it produces, and figures were verified against the underlying data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All leaderboards, ranking tables, fitted comparison sets, extraction responses, model catalogs and per-board method notes are released with the paper; the code is provided as an anonymized archive in the supplementary material. The extraction prompt is reproduced verbatim in Appendix B, the normalization rules in Appendix C, and the identifiability restriction in Algorithm 1 with per-board effects in Appendix D. Appendix F reproduces three views that ship with the pathology board; every other figure and table in the paper is produced by a single script that reads only released artifacts.

REFERENCES

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, B. Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram S. Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, T. Arbel, Christopher Joseph Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/BIOMET/39.3-4.324.

Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433, 2012. doi: 10.1214/12-STS396.

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

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning (ICML)*, pp. 8359–8388, 2024.

Gregory R. Conner and Christopher P. Grant. An extension of Zermelo's model for ranking by paired comparisons. *European Journal of Applied Mathematics*, 11(3):225–247, 2000. doi: 10.1017/S0956792500004150.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024. doi: 10.1038/s41467-024-45563-x.

Roger R. Davidson. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

R. Dittrich, R. Hatzinger, and W. Katzenbeisser. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):511–525, 1998. doi: 10.1111/1467-9876.00125.

Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, 1997. doi: 10.1136/BMJ.315.7109.629.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Julien Fageot, Sadegh Farhadkhani, Lê-Nguyên Hoang, and Oscar Villemaud. Generalized Bradley-Terry models for score estimation from paired comparisons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20379–20386, 2024. doi: 10.1609/aaai.v38i18.30020.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. doi: 10.1093/biomet/80.1.27.

L. R. Ford. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8, Part 2):28–33, 1957. doi: 10.1080/00029890.1957.11989117.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359 (6379):eaao0185, 2018. doi: 10.1126/science.aao0185.

Gene V. Glass. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10): 3–8, 1976. doi: 10.3102/0013189X005010003.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Julian P. T. Higgins and Simon G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002. doi: 10.1002/SIM.1186.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5203–5213, 2019. doi: 10.18653/v1/P19-1513.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 221–233, 1967.

David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *arXiv preprint arXiv:1511.08060*, 2015.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004. doi: 10.1214/AOS/1079120141.

John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. Meta-research: Evaluation and improvement of research methods and practices. *PLOS Biology*, 13(10):e1002264, 2015. doi: 10.1371/journal.pbio.1002264.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnić. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8580–8594, 2020. doi: 10.18653/v1/2020.emnlp-main.692.

Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1):71–82, 2021. doi: 10.1093/biomet/asaa052.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, D. Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan A. Chi, Sang Michael Xie, Shibani Santurkar, S. Ganguli, Tatsunori B. Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian L. Hie, Zhong-Kai Zhu, Wen-Ting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Y. Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, G. Jaume, Igor Odintsov, L. Le, Georg K. Gerber, A. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30 (3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7.

Benjamin Marie, Atsushi Fujita, and Raphaël Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7297–7306, 2021. doi: 10.18653/v1/2021.acl-long.566.

Mark E. J. Newman. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25, 2023.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Y. Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71, 2021. doi: 10.1136/bmj.n71.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily L. Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

A. Springall. Response surface fitting using a generalization of the Bradley-Terry paired comparison model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(1):59–68, 1973. doi: 10.2307/2346303.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, A. Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, A. Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, A. Santilli, Andreas Stuhlmüller, Andrew M. Dai, A. La, A. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, A. Tabassum, Arul Menezes, Arun Kirubarajan, A. Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. R. Roberts, B. S. Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, B. Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C. Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Christopher Callison-Burch, C. Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, D. Kilman, Dan Roth, D. Freeman, Daniel Khashabi, Daniel Levy, D. M. González, Danielle Perszyk, Danny Hernandez, Dan-Qi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, E. Shutova, E. D. Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, E. Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, E. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, H. Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, J. F. Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, J. Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, J. Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jia-Cheng Xu, Jiaming Song, Jillian Tang, Jane W. Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jorg Frohberg,

Jos Rozen, J. Hernández-Orallo, Joseph Boudeman, J. Guerr, Joseph Jones, Joshua B. Tenenbaum, Josh Rule, Joyce Chua, Kamil Kanclerz, K. Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, K. Markert, Kaustubh D. Dhole, Kevin Gimpel, K. Omondi, K. Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M. Ramírez-Quintana, M. Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, M. Schubert, Medina Baitemirova, Melody Arnaud, M. McElrath, Michael A. Yee, M. Cohen, Michael Gu, Michael I. Ivanitskiy, Michael Starritt, M. Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, M. Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T. MukundVarma, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, N. Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, P. Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, P. Eckersley, Phu Mon Htut, P. Hwang, P. Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphael Milliere, Rhythm Garg, Richard L. Barnes, R. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, R. Salakhutdinov, Ryan A. Chi, Ryan Lee, Ryan Stovall, R. Teehan, Rylan Yang, Sahib Singh, Saif Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Samuel Wiseman, Samuel Gruetter, Samuel R. Bowman, S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shi-Xiang-Sha-Ne Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, S. Melzi, Siva Reddy, S. Makini, Soo-Hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, S. Dehaene, Stefan Divic, S. Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, T. Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, T. Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. Kornev, T. Tunduny, Tobias Gerstenberg, T. Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, W. Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu-Fang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zi-Jian Wang, Zi-Jie J. Wang, Zi-Rui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2): 146–160, 1972. doi: 10.1137/0201010.

Heather Turner and David Firth. Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 48(9):1–21, 2012. doi: 10.18637/jss.v048.i09.

U.S. National Science Foundation. Science of Science: Discovery, Communication and Impact (SoS:DCI). `https://www.nsf.gov/funding/opportunities/sosdci-science-science-discovery-communication-impact`, 2026. Funding program; accessed September 2026.

John T. Whelan. Prior distributions for the Bradley-Terry model of paired comparisons. *arXiv preprint arXiv:1712.05311*, 2017.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. doi: 10.2307/1912934.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/c7sc02664a.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, C. Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu-Hsin Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chun-yue Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, R. Weerasinghe, Bill J. Wright, Ari Robicsek, B. Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Ting Yan. Ranking in the generalized Bradley–Terry models when the strong connection condition fails. *Communications in Statistics – Theory and Methods*, 45(2):340–353, 2016. doi: 10.1080/03610926.2013.809114.

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 (Datasets and Benchmarks Track)*, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

A Search Workflow Details

The five inputs. The Phase 1 skill accepts exactly `topic`, `model_groups` (one or more object groups), `input` (the modality the objects consume), `output` (one or more output types) and `time` (an ISO cutoff or an interval). Everything else has a default. Inputs 2–4 are composed into the target-capability sentence

> *The object directly takes* `<input>` *as input and produces* `<output...>`*, and belongs to the object group(s):* `<model_groups...>`.

which is written into the frozen configuration and is the sole criterion for condition 2. Optional refinements (seed terms, exclusion terms, boundary cases, track and source selection) do not change this contract. The skill refuses to run with a missing input, a non-ISO or inverted time window, or an input that contradicts the declared object groups.

Run directory and deliverables. Initialization writes `input_config.yaml`, `scope_snapshot.yaml`, `benchmark_spec_snapshot.yaml`, a `search_plan.json` skeleton, `state.json` and per-file SHA-256 digests; a supplied model list is snapshotted and hashed alongside them. Retrieval adds `source_responses/` (one file per source, query and page) and `source_manifest.json`. Gating adds `candidates.csv`, `strict_review_pool.csv` and the three decision tables (`strict_accepted.csv`, `strict_borderline.csv`, `strict_excluded.csv`), plus `graded_report.md` and `graded_report.jsonl` and a one-page `review_request.md` stating the search boundary, the verbatim target capability, whether pagination was exhausted per source, and which files to read first. Each candidate record carries `new_object_evidence`, `target_topic_evidence`, `primary_record_evidence`, `condition_results`, `decision`, `decision_reason`, `judge_only_risk`, `specialist_primary_risk` and `group_ids`. Human decisions made during review are written back into `input_config.yaml` as `manual_inclusions`, `hard_exclusions` or `borderline` rather than edited into the final CSV, so that a re-run does not lose them.

The three strict conditions. **(1) New public object.** The models track requires a paper, official project, repository or model card to propose or release a new model, method, version or public record; the benchmarks track requires the creation or release of a new benchmark, evaluation suite, testbed, protocol, split or test set. Appearing as a baseline, being mentioned, or merely comparing models on an existing benchmark does not count. **(2) Target topic capability.** The object must directly handle or evaluate the capability, task, input, output or modality defined for the configured domain, with verifiable task, result or protocol evidence. **(3) Primary object record.** A trustworthy

first-hand public record must identify the model or benchmark and its domain capability. A paper-only record on arXiv or in a formal venue passes; a checkpoint, complete dataset or repository is not required at Phase 1, and missing fields are written `unknown`. An object that is only a judge, grader, annotator, data or trajectory generator, reward or scoring model, or other auxiliary component does not pass.

Candidates are graded A (directly defines or releases the target object with complete evidence for all three conditions), B (highly relevant but missing one key piece of evidence, or ambiguous at the boundary) or C (background, method transfer or weak relevance, requiring human judgment). Grades express screening confidence and relevance for this task only; they are not statements about paper quality or scientific validity.

B THE EXTRACTION PROMPT

The prompt below is the exact text sent with every PDF, reproduced from the released `paper_extraction_prompt.txt`. The topic-scope paragraph is the only part that changes between boards; the version shown is the pathology foundation model board.

```
Analyze the uploaded PDF and output JSON only.
First classify the paper:  model_paper (introduces, trains,
releases or primarily presents a model/system and evaluates
it); benchmark (primarily introduces or evaluates a
benchmark, dataset, evaluation suite, protocol or
leaderboard rather than a single focal model);
model_and_benchmark (both); other (neither -- return no
records).
Topic scope (strict on input, open on output).  The
evaluated system must take a histopathology image as input
(a whole-slide image, a WSI tile/patch, or a
tissue-microarray core, H&E or IHC stained), alone or with
explicitly documented image-derived representations.  The
output is not restricted:  any evaluated task or output type
counts.  Include all eligible downstream, fine-tuned,
linear-probe, few-shot or zero-shot evaluations, including
external-cohort generalization.  Exclude systems whose
primary input is natural-language text, omics sequences,
radiology images, molecular graphs or other
non-histopathology modalities unless the paper clearly
evaluates the histopathology-image input.
Model-paper and mixed-paper rules.  (1) When paper_type is
model_paper or model_and_benchmark, identify the exact
proposed/focal model name from the PDF (title, abstract,
method, tables or captions) and put it in the top-level
proposed_model_name field.  Never infer this name from the
filename.  (2) Exclude evaluations of that proposed model,
including spelling/case variants, dated or size variants,
reasoning/configuration variants, and checkpoints or
fine-tuned versions belonging to the same proposed release.
Treat Ours, Our model and This work as the proposed model
when they appear in an evaluation table.  (3) For
model_paper, extract evaluations for all other valid
deep-learning models reported in eligible benchmark blocks.
For model_and_benchmark, extract every valid evaluation in
the newly introduced benchmark(s) and any other eligible
benchmark blocks, while still excluding the proposed model
under rules 1-2.  Exclude ablations and component-removed
variants.
Benchmark-paper rules.  (4) When paper_type is benchmark,
set proposed_model_name to an empty string and record every
valid evaluation of every reported deep-learning model,
including baselines and models used by the benchmark
```

```
authors.  Do not omit a row merely because it is not a
headline result.  Exclude only non-model methods and
explicit ablation/component-removed variants.

Common extraction rules.  (5) Record all eligible
benchmarks, datasets, subsets, splits, tasks and metrics; do
not cherry-pick.  (6) Keep raw reported values exactly.  Do
not perform pairwise comparisons, ranking, winner inference,
unit conversion, or estimation from plots.  A plot value may
be used only when an exact numeric value is explicitly
printed and legible.  (7) For every metric, set
higher_is_better to true or false using the paper's
definition or direction marker.  If the direction cannot be
established, use null.  (8) A valid evaluation must have an
identifiable model, benchmark/task context, metric, and an
explicitly reported numeric score.

Output constraints.  Output JSON only; no Markdown fences or
explanatory text.  Use exactly model_paper, benchmark,
model_and_benchmark or other for paper_type.  If no valid
evaluation exists, return an empty records array.
```

The required schema is an object with `paper_type`, `proposed_model_name`, and a `records` array whose elements carry `paper_title`, `benchmark`, `task`, `dataset`, `model_name` and a `scores` array of `{metric, value, higher_is_better}`. Responses that fail schema validation, or that omit `proposed_model_name` on a model paper, are rejected and retried; a PDF whose response never validates is recorded as failed rather than silently contributing nothing. Transient failures are retried with backoff, while deterministic failures (missing route, unauthorized, bad request) stop immediately.

C Normalization Rules

Score parsing. One regular expression accepts an optional leading sign, a decimal or scientific mantissa, an optional percent sign, and an optional ± (also +/−, +−) dispersion term. Mean and dispersion are stored separately; only the mean is used to form battles.

Metric direction precedence. (1) The `higher_is_better` value returned by the extractor, when present. (2) An explicit override table for metric names the keyword rules would read backwards: lower-is-better for `fnr`, `fpr`, `error rate`, `validation loss`, `training loss`, `params (m)`, `flops (g)`, `model size (mb)`, `inference time (ms)`; higher-is-better for `tnr`, `tpr`, `sensitivity (%)`, `specificity (%)`. (3) Context-sensitive regular expressions evaluated against the benchmark, task and dataset fields, which catch adversarial and safety quantities: attack success probability or rate, successful attacks in attempts or scenarios, pre-attack target-choice rate, and false/missed/harmful/unmatched/unparseable alarm, safe, reference, request or response rates. These run *before* the metric-only rules so that "attack success rate" is not read as a desirable success rate. (4) Keyword lists over the metric name. Lower-is-better keywords include `error`, `loss`, `mae`, `mse`, `rmse`, `ece`, `brier`, `fid`, `lpips`, `perplexity`, `wer`, `cer`, `latency`, `runtime`, `hallucination`, `cost` and `distance`; higher-is-better keywords include `accuracy`, `auc`, `auroc`, `auprc`, `f1`, `iou`, `map`, `mrr`, `ndcg`, `bleu`, `rouge`, `cider`, `precision`, `recall`, `correlation`, `success rate` and `win rate`.

Unordered metrics. Metrics that report a value rather than a score have no better-or-worse direction. Their rows are dropped before any battle is formed rather than ranked under a guessed sign.

Model identity. Canonicalization is a list of regular expressions tried in order, and the first match names the row, so the order is part of the specification rather than an implementation detail: name guards, then domain foundation models, then self-supervised pre-training methods, then generic encoders, then MIL aggregators, then everything else.

Algorithm 1 Identifiable fitting set

Require: battles B, threshold τ, optional allowlist A
canonicalize names; drop unresolved and self-battles
drop models with corpus count $< \tau$; keep battles inside A
repeat
  $V_{\text{prev}} \leftarrow \text{models}(B)$
  repeat ▷ evidence k-core
    drop models with count$(B) < \tau$ and their battles
  until no model is below τ
  $C \leftarrow$ largest strongly connected component of B
  restrict B to battles with both endpoints in C
until $\text{models}(B) = V_{\text{prev}}$
recompute pair-frequency weights on B
assert $|\text{components}(B)| = 1$ and $\text{mincount} \geq \tau$

Name guards are anchored rules for labels a later, unanchored rule would misread, and two kinds occur. A proposed system can carry a foundation-model token inside its own name — `ACM-UNI`, `CellViT-UNI`, `MaskFormer-UNI`, `UniMiSS` — and without a guard the `UNI` rule would claim those rows and credit UNI with another model's results. And a slide-level foundation model is often labelled by the tile encoder it runs on, as in `CHIEF (CTransPath tile encoder)` or `TITAN (CONCH v1.5)`; such a row reports the slide encoder, so the head of the label wins even though encoder rules are otherwise free to match anywhere.

Within a family the more specific checkpoint always precedes the shorter prefix, so `UNI2-h` is not absorbed by the `UNI` rule and two separate releases do not merge into one player. Encoder families match anywhere in the label while aggregator families are anchored at its head, and the two together implement the encoder-wins attribution: `AB-MIL (ResNet-50)` resolves to ResNet-50 and `CLAM-SB+CTransPath` to CTransPath, because the encoder rule is both earlier in the order and free to match away from the head; `UNI + ABMIL` and `HistGen w/ PLIP` resolve to UNI and PLIP for the same reason.

Drop rules remove labels that name no single checkpoint, because a retained one would become a fictitious player accumulating battles from papers with nothing in common. They cover cross-paper aggregates (`mean-of-3`, `average-over`, `backbone with MIL aggregators averaged`) and A-vs-B labels; generic architecture names (`CNN`, `Custom CNN`), which denote a different hand-built network in every paper; positions in one paper's ablation (`Baseline`, `Early fusion`, `Unimodal`, `WSI Only`), which describe a configuration of that paper's own system rather than a released model; and author or challenge-team handles (`Coudray et al.`), which name a group rather than a checkpoint.

D IDENTIFIABILITY AUDIT

The Newton step. The objective of (2) has gradient and Hessian

$$g(\beta) = \sum_r w_r \,(\sigma_r - y_r)\, x_r, \qquad H(\beta) = \sum_r w_r \,\sigma_r(1-\sigma_r)\, x_r x_r^\top, \qquad \sigma_r \equiv \sigma(\beta^\top x_r), \tag{5}$$

and we iterate $\beta \leftarrow \beta - t\, H^+ g$ under the reference coding, with H^+ the Moore–Penrose pseudo-inverse, so a direction the design leaves flat receives no update instead of an unbounded one, and $t \in \{1, \frac{1}{2}, \frac{1}{4}, \dots\}$ chosen by backtracking until the objective decreases. The fit is accepted at $\|g\|_\infty < 10^{-5}$, then centered onto $\sum_i \beta_i = 0$, and (3) is evaluated at $\hat{\beta}$ and carried through the same centering.

Table 2 reports per-board precision diagnostics together with whether the released fitting set satisfies Ford's condition. We checked every board: all fourteen comparison graphs are strongly connected as released.

The clearest symptom of a violation is a model with an observed win rate of exactly 0 or 1, which has no outgoing (or no incoming) edge and is therefore its own strongly connected component; its fitted score is wherever the Newton iteration stopped rather than an estimate. An earlier fit of the

Vision (2026) board, run before the restriction was enforced, failed in exactly this way: one model had won none of its 50 comparisons, its parameter diverged, and the fit returned a score of -4458 — the point where the iteration stopped, not an estimate. Refitting that board under the restriction removed the model, reducing the board from 62 to 48 models. No model on any currently released board has an observed win rate of 0 or 1. The audit script that produces this table is released with the paper.

What a penalized refit would involve. Section 4.2 declines the estimators that fit outside the identifiable subgraph; this records what they are and what adopting one would cost, since it is the obvious extension of these boards. Four families relax Ford's condition. A vanishing singular perturbation of the win matrix produces a ranking for *any* outcome matrix (Conner & Grant, 2000), which Yan (2016) extends to generalized Bradley–Terry models with ties and order effects, with necessary and sufficient conditions for the perturbed estimate to exist and be unique. A Gaussian prior on the strengths makes the posterior mode exist, be unique and be computable on every comparison graph, for a whole exponential family of comparison models (Fageot et al., 2024). A Jeffreys-prior penalty buys the same finiteness with no tuning constant and a bias reduction besides (Firth, 1993; Kosmidis & Firth, 2021). And structured variants go further than existence, scoring models with no usable comparisons at all by making strength a function of model covariates (Springall, 1973; Turner & Firth, 2012), or a random effect that shrinks toward the board mean while modelling the dependence between comparisons directly (Cattelan, 2012) — the dependence our HC1 intervals currently correct for after the fact. A parallel line leaves the strengths alone and lets the comparison depend on covariates of the *judge* (Dittrich et al., 1998); on a literature board the judge is the source paper, so that is the natural place to model a self-report or venue effect rather than exclude it.

The cheapest version adds a Gaussian prior on the strengths, which is two terms in the existing Newton step and leaves the estimate on the identifiable subgraph unchanged as the penalty vanishes, so the restricted fit is recovered as a limiting case rather than contradicted (Fageot et al., 2024). Three things do not come for free. The penalty must be applied to the centered strengths and not to the reference-coded design, or it shrinks the board toward whichever model was picked as the reference instead of toward the mean — an invariance requirement Whelan (2017) states explicitly as a criterion for choosing a prior at all. Its constant has to be chosen by cross-validation grouped at the level of the source paper, because the $\binom{k}{2}$ battles a single table emits are not exchangeable and an ungrouped split would report a badly optimistic penalty. And the HC1 intervals of §4 would need the penalized Hessian as their bread, or they would be narrower than the fit supports. The scores of recovered models would then have to be flagged as prior-driven rather than data-identified, which is the reporting change that matters most: the point of the restriction is that no such flag is currently needed on any released row.

Sparsity and the shape of literature graphs. The restriction bites in proportion to how sparse a board is, and sparsity scales with the population it admits. Pooled across the 14 boards (Figure 4), observed pairs grow as $M^{1.35}$ against the M^2 of a complete graph, so density falls as $M^{-0.65}$. Per model the regimes largely separate: a listed board observes a median of 41 distinct opponents per model against 15 for an open board, over ranges of 19–87 and 9–20 that overlap only at their edges; that count is what a sparse board is short of. Open literature graphs are not merely sparse but *hub-organized*, built around a few heavily cited reference models — the structure that makes Ford's condition bite, and also what keeps those boards connected. Spread differs widely too, driven by population heterogeneity rather than domain: the interquartile range of fitted scores runs from 264 Arena points (Reasoning & Knowledge) to 1,225 (Math, sub-2B to frontier), so a wide board spans more tiers rather than discriminating better.

E One Board in Full

Every released board ships the two views below, and the main text shows only the top of two boards. One complete board is reproduced here so that what a literature-derived leaderboard looks like below its leaders — where the intervals are wide, the ordering is not resolved, and most of the population lives — is visible rather than asserted. Spatial transcriptomics is used because it is the smallest open-population board (52 models) and therefore the one that fits at a legible row pitch; the corresponding figures for the other thirteen are in the release.

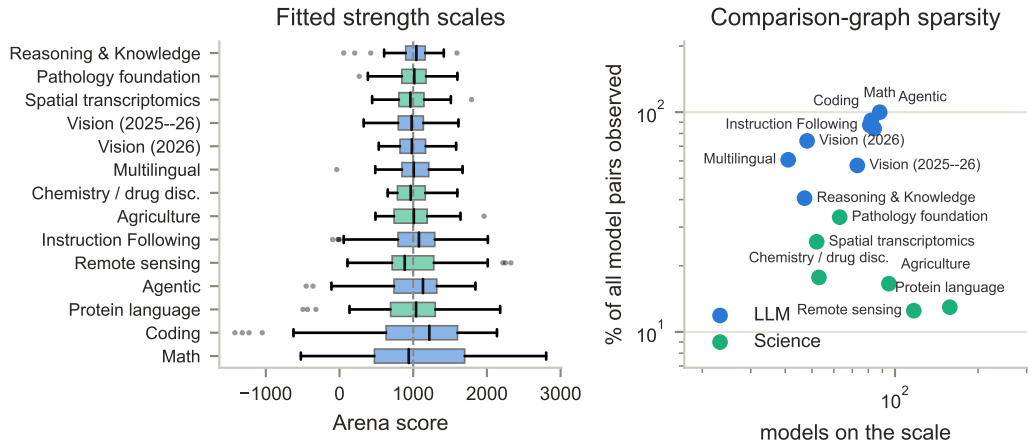


Figure 4: Board-level structure. **Left:** distribution of fitted Arena scores per board, ordered by interquartile spread; the dashed line is the board mean of 1000. Spread tracks how many capability tiers a board's population spans, not the domain. **Right:** comparison-graph density against board size; boards fitted against a curated list are near-complete, open-population boards are hub-organized.

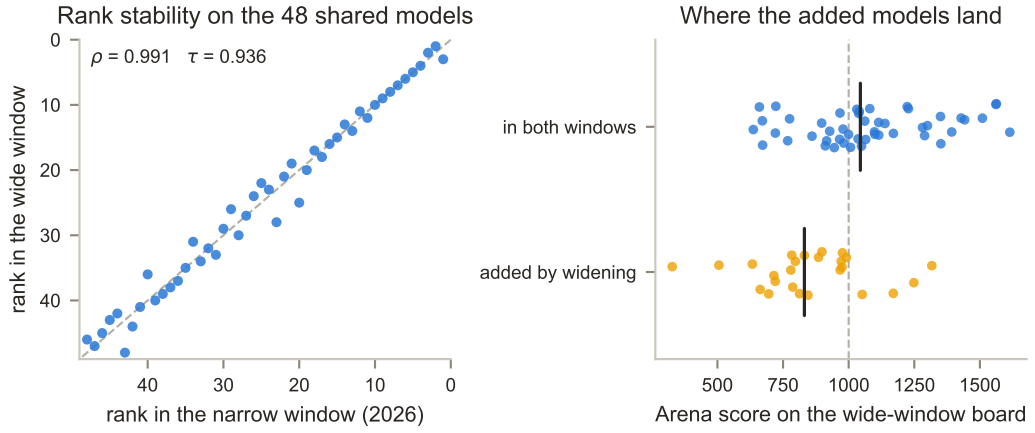


Figure 5: Two nested collection windows over the vision track (§6.3). **Left:** rank on the 2026-only board against rank on the 2025–2026 board, for the 48 models both contain, with ranks recomputed inside the shared set so the added models cannot inflate the agreement. **Right:** where the 25 models the wider window adds sit on the wider board's scale; vertical bars are group medians and the dashed line is the board mean of 1000. Widening the window extends the board downward without reordering it.

Figure 6 is the complete ranking. The leaders separate, the long middle does not, and the visible width of the intervals is the honest content of the board: neighbouring rows in the middle third are ordered by the point estimate and not distinguished by the evidence. Figure 7 plots the same models against the evidence behind them, with marker area proportional to precision. It shows the two failure modes a single ranked list hides — a high score resting on few comparisons, and a heavily compared model whose interval is still wide because its comparisons disagree — and it is the view we expect a reader checking a specific model to want.

Table 2: Per-board diagnostics. *CI* is the median 95% interval width in Arena points; *IQR* is the interquartile spread of fitted scores; *Separated* is the fraction of model pairs whose intervals do not overlap; *Unique rank* is the fraction of models whose rank spread is a single position; *Pool* is `list` when the board was restricted to the curated model list and `open` when it admits every model the literature named; *Ford* indicates whether the released fitting set satisfies Ford's strong-connectivity condition.

Leaderboard	Models	Ties	CI	IQR	Separated	Unique rank	Pool	Ford
General LLM / VLM capabilities								
Reasoning & Knowledge	47	453	107	264	73.4%	0.0%	list	yes
Math	88	175	182	1225	87.3%	0.0%	list	yes
Coding	82	707	163	969	87.8%	0.0%	list	yes
Agentic	84	3,383	92	580	86.3%	0.0%	list	yes
Instruction Following	81	117	106	498	84.8%	0.0%	list	yes
Multilingual	41	150	154	363	73.9%	2.4%	list	yes
Vision (2025–26)	73	1,982	80	341	79.8%	1.4%	list	yes
Vision (2026)	48	1,040	104	346	76.8%	0.0%	list	yes
Scientific domains								
Pathology foundation	63	1,031	102	328	73.5%	0.0%	list	yes
Protein language	158	1,245	224	606	66.0%	0.0%	open	yes
Remote sensing	117	65	244	565	67.6%	0.0%	open	yes
Agriculture	95	375	198	450	64.3%	1.1%	open	yes
Chemistry / drug disc.	53	54	298	378	34.6%	0.0%	open	yes
Spatial transcriptomics	52	86	139	341	71.5%	0.0%	open	yes

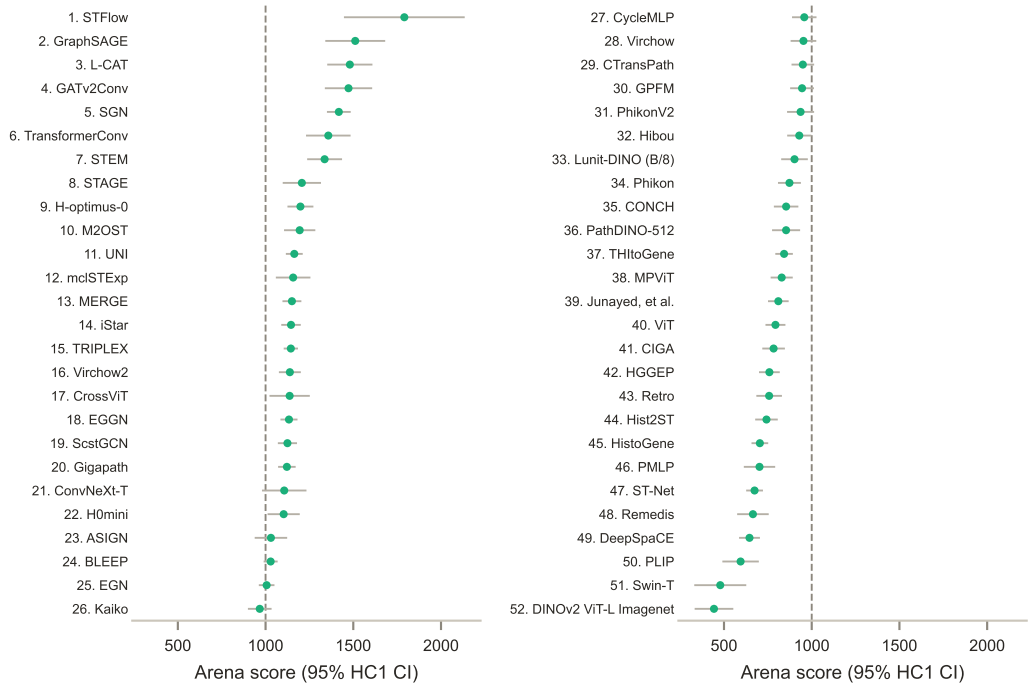


Figure 6: The spatial transcriptomics board in full: all 52 fitted models with 95% HC1 intervals, split into two columns on a shared score axis. The dashed line is the board mean of 1000.

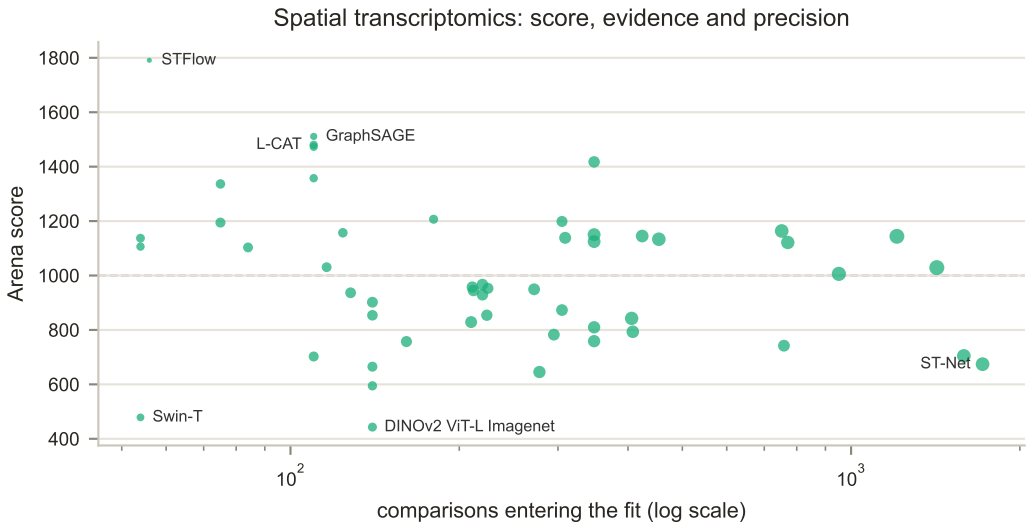


Figure 7: Score against evidence for the same board. Marker area is proportional to $1/\text{interval width}$, so a large dot is a precisely placed score. Labels mark the three highest, the two lowest and the most-compared model.

F THE COMPARISON GRAPH BEHIND A BOARD

A ranked list hides the object the ranking is computed from. The three views below show it directly for the pathology foundation model board, the largest listed scientific board, and they are reproduced as released rather than redrawn: they ship with that board's pairwise dashboard, which is where a reader checks a specific pairing. All three are defined over the 63 models of the released fitting set and count a battle only when both endpoints are on it, so they describe exactly the 51,269 comparisons §5 reports. The display then applies the $\tau = 50$ floor of §4.2 to *non-tied* matches alone, which is a slightly stricter test than the fit itself applies: a tie is evidence that a model took part in a comparison, so the fit counts it, but it lands in no cell of a win matrix, so the display cannot. Exactly one model fails the stricter test. PathGen-CLIP-B enters the fit with 54 comparisons and is ranked 28th on the released board, but five of them are ties, leaving 49 non-tied matches — one short of the floor. It is removed in a single round together with all 54 of its comparisons, and no further model then falls below the threshold. The three figures therefore show 62 models over 50,189 non-tied battles with 1,026 ties set aside, and the counts reconcile as $50{,}189 + 1{,}026 + 54 = 51{,}269$. Nothing is refitted: the scores and intervals reported everywhere else remain those of the 63-model fit, and PathGen-CLIP-B keeps its place on the board of Appendix G.

Figure 8 is the empirical win matrix — the quantity the Bradley–Terry fit of (2) summarizes into one number per model — and Figure 9 is the evidence behind each of its cells. Read together they are the argument for the machinery of §4.2. White is the dominant colour in both: most pairs are never compared, which is the sparsity measured in Appendix D, and a rank cannot rest on a cell that does not exist. The coloured mass is organised in columns rather than spread evenly, because a few reference models — UNI, CONCH, CTransPath, Prov-GigaPath — are the opponent almost everyone reports against; the battle-count view shows the same columns carrying thousands of battles while most cells carry tens. That is the hub structure a literature graph has and a sampled arena does not, and it is what makes Ford's condition a real constraint rather than a formality: a model compared only against one hub, and only in one direction, is exactly the model the fixed-point loop removes.

Figure 10 is the ordering these two views imply on their own, before any model is fitted. It is worth comparing with the released board (Appendix G): an average win rate treats a model that beat two weak opponents once and a model that beat forty strong ones as interchangeable, which is why PathGen-CLIP-L leads at 97.5% on 79 non-tied matches. The Bradley–Terry scale exists to discount exactly that, and the interval attached to each score is what says whether the discount was enough.

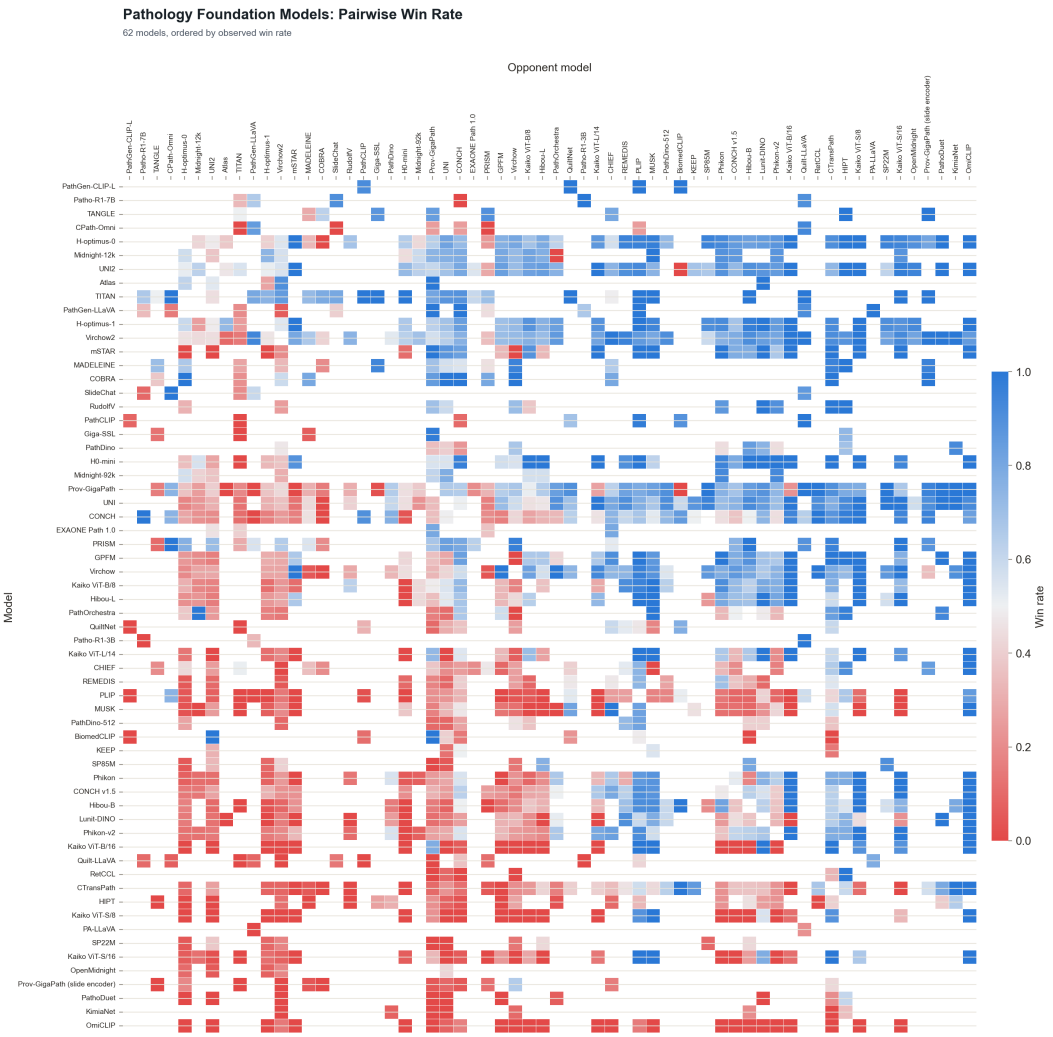


Figure 8: Empirical pairwise win rate on the pathology foundation model board, row model against column model, ordered by average win rate. Blue is a row win, red a row loss, white an unobserved pair. Most of the matrix is unobserved, and the observed mass concentrates in the columns of a few heavily cited reference models.

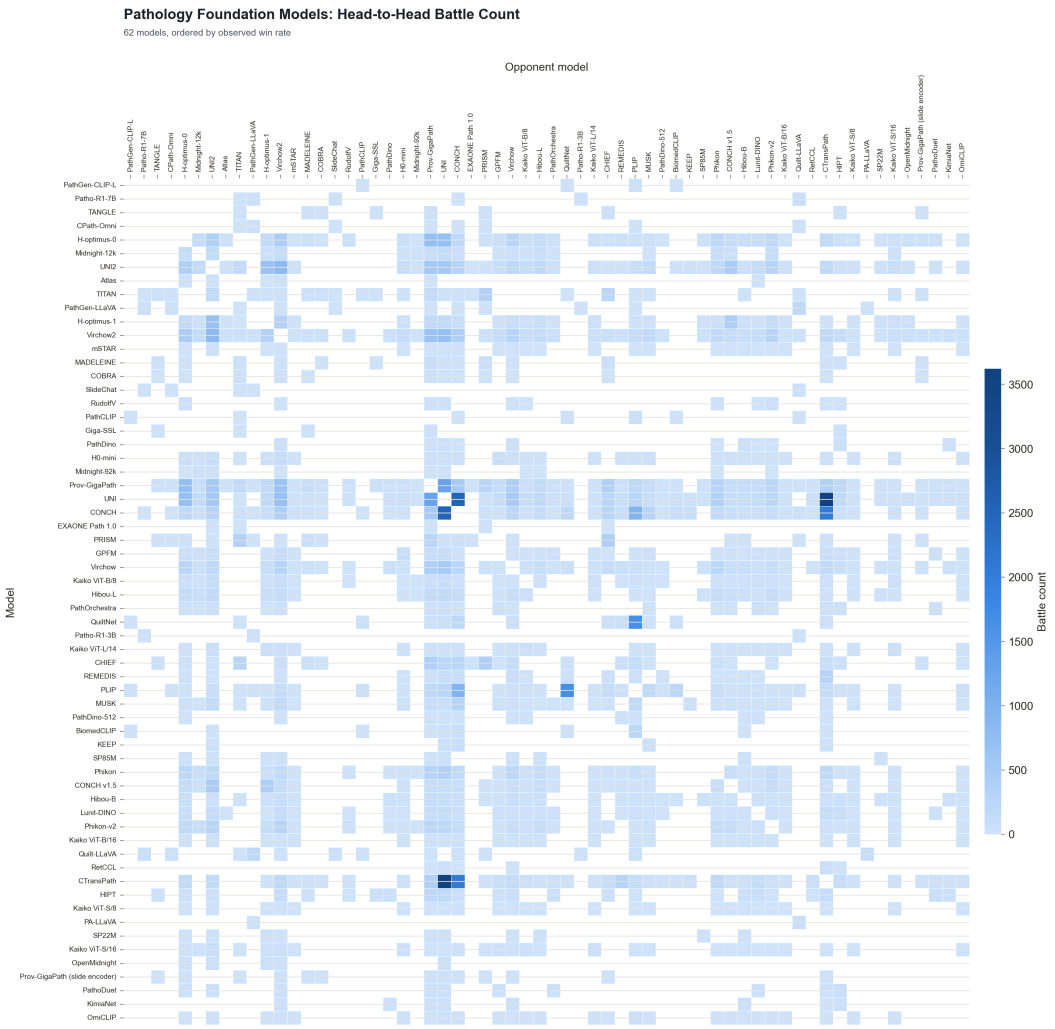


Figure 9: Head-to-head battle counts for the same pairs and the same ordering. Colour is the number of comparisons behind each cell of Figure 8: a handful of pairings carry thousands of battles while most carry tens, so a uniform reading of the win matrix would be dominated by a few repeatedly reported pairs — the imbalance the weights of (2) correct.

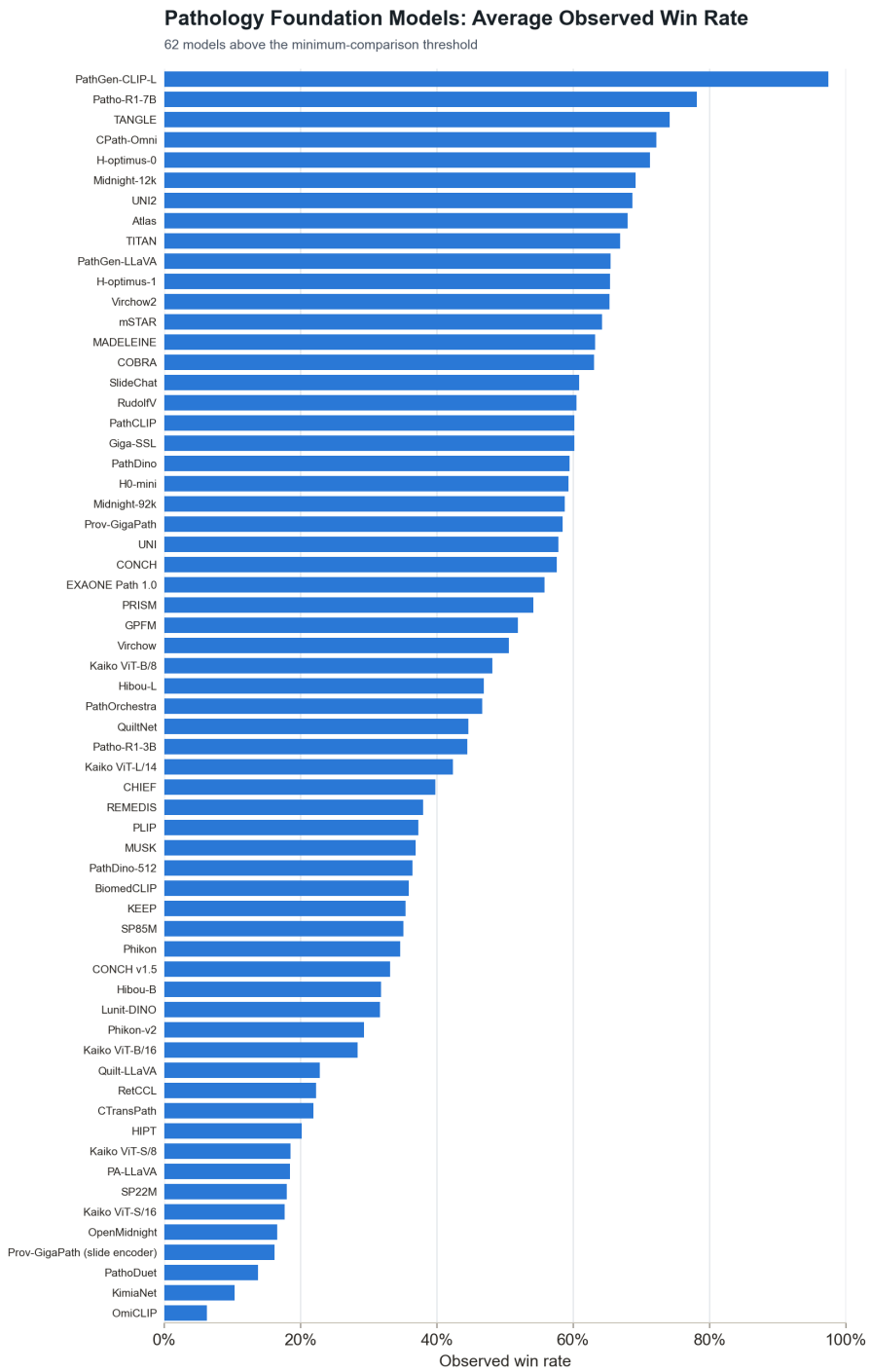


Figure 10: Average observed win rate per model, the row means of Figure 8. This is the ranking the raw comparisons give without a model: it ignores who each model played and how often, so a short record against weak opponents outranks a long record against strong ones.

G LEADERBOARD EXTRACTS

The top ten entries of every released board follow. Scores are relative strengths on the Arena scale, not accuracy percentages, and *Votes* is the number of fitted comparisons the model took part in. Complete tables, including rank spreads, are in the release.

Reasoning & Knowledge

#	Model	Score	95% CI	Votes
1	Claude Mythos 5	1593	1378–1809	522
2	GPT-5.4 Pro	1415	1297–1533	297
3	GPT-5.6 Sol	1384	1302–1466	1,094
4	Claude Fable 5	1361	1296–1426	351
5	Claude Opus 5	1340	1286–1393	795
6	Kimi K3	1314	1189–1438	195
7	Claude Opus 4.8	1290	1252–1329	899
8	GPT-5.5	1286	1241–1331	3,628
9	Gemini 3.1 Pro	1215	1188–1241	4,647
10	Claude Opus 4.7	1214	1147–1281	1,093

Math

#	Model	Score	95% CI	Votes
1	Claude Fable 5	2804	2594–3015	227
2	GPT-5.6 Sol	2534	2391–2677	222
3	Claude Opus 5	2469	2293–2645	202
4	Muse Spark 1.1	2388	2227–2549	174
5	Kimi K3	2315	2174–2455	199
6	GPT-5.5	2307	2217–2398	623
7	Claude Opus 4.8	2122	2017–2228	378
8	GPT-5.6 Terra	2103	1999–2207	174
9	Gemini 3.1 Pro	2082	1968–2196	1,027
10	Claude Opus 4.7	2080	1994–2167	345

Coding

#	Model	Score	95% CI	Votes
1	Claude Mythos 5	2136	2026–2247	197
2	Claude Fable 5	2065	2009–2121	1,120
3	GPT-5.6 Sol	2029	1971–2086	2,158
4	Claude Opus 5	2019	1943–2094	1,645
5	Kimi K3	1986	1936–2036	1,424
6	GPT-5.6 Terra	1974	1907–2041	440
7	Grok 4.5	1973	1892–2054	313
8	Muse Spark 1.1	1906	1806–2007	160
9	Claude Opus 4.7	1905	1842–1968	2,487
10	GPT-5.5	1885	1824–1947	3,351

Agentic

#	Model	Score	95% CI	Votes
1	Claude Opus 5	1844	1774–1914	503
2	Claude Fable 5	1729	1683–1776	1,275
3	Claude Mythos 5	1638	1583–1692	1,424
4	Kimi K3	1584	1551–1616	1,401
5	Claude Sonnet 5	1557	1520–1595	1,088
6	GPT-5.6 Sol	1546	1487–1605	1,758
7	Qwen3.8 Max	1524	1480–1569	694
8	Grok 4.5	1517	1477–1556	564
9	GPT-5.6 Luna	1481	1432–1529	351
10	GPT-5.6 Terra	1478	1418–1538	774

Instruction Following

#	Model	Score	95% CI	Votes
1	Claude Opus 5	2013	1863–2163	253
2	Qwen3.8 Max	1822	1716–1927	223
3	Kimi K3	1808	1729–1888	338
4	Claude Mythos 5	1699	1522–1875	205
5	Grok 4.5	1623	1568–1678	338
6	Claude Fable 5	1583	1513–1654	561
7	Claude Sonnet 5	1566	1512–1619	311
8	GPT-5.5	1505	1456–1553	863
9	GPT-5.6 Sol	1489	1423–1555	564
10	GPT-5.6 Luna	1476	1423–1529	273

Multilingual

#	Model	Score	95% CI	Votes
1	Claude Fable 5	1670	1582–1758	206
2	Claude Mythos 5	1640	1524–1757	122
3	Qwen3.6 Plus	1423	1203–1643	67
4	Gemini 3.5 Flash	1421	1373–1468	340
5	Sarvam 30B	1396	1284–1507	52
6	Gemini 3.1 Pro	1382	1308–1456	840
7	GPT-5.5	1363	1302–1424	329
8	Claude Opus 4.8	1347	1298–1395	350
9	Claude Opus 4.6	1316	1264–1368	887
10	Claude Opus 4.7	1270	1221–1319	279

Vision (2025–26)

#	Model	Score	95% CI	Votes
1	Claude Mythos 5	1614	1493–1735	232
2	GPT-5.6 Sol	1563	1505–1620	803
3	Claude Opus 5	1561	1466–1656	201
4	Kimi K3	1510	1416–1603	206
5	Claude Fable 5	1442	1393–1490	473
6	Seed-2.1-Pro	1428	1360–1497	154
7	GPT-5.4 Pro	1393	1329–1456	323
8	GPT-5.6 Terra	1351	1253–1449	201
9	Gemini 3.5 Flash	1350	1300–1400	864
10	GPT-5.2 Pro	1317	1270–1364	360

Vision (2026)

#	Model	Score	95% CI	Votes
1	Claude Opus 5	1583	1480–1685	186
2	Claude Mythos 5	1565	1450–1679	230
3	GPT-5.6 Sol	1565	1504–1625	788
4	Kimi K3	1454	1362–1546	206
5	Claude Fable 5	1385	1335–1435	473
6	Seed-2.1-Pro	1374	1303–1444	154
7	GPT-5.4 Pro	1338	1277–1399	218
8	GPT-5.6 Terra	1316	1204–1428	186
9	Gemini 3.5 Flash	1288	1235–1341	856
10	Gemini 3.1 Pro	1261	1233–1289	8,109

Pathology foundation

#	Model	Score	95% CI	Votes
1	PathGen-CLIP-L	1600	1376–1823	94
2	COBRA	1431	1348–1514	142
3	TITAN	1409	1362–1456	1,220
4	TANGLE	1387	1312–1463	200
5	Atlas	1355	1275–1436	79
6	H-optimus-1	1308	1275–1341	2,839
7	PRISM	1281	1232–1330	1,355
8	MADELEINE	1269	1220–1318	499
9	H0-mini	1263	1229–1297	794
10	Midnight-12k	1257	1215–1299	1,262

Protein language

#	Model	Score	95% CI	Votes
1	AIDO-Protein-RAG	2180	2030–2330	242
2	VenusREM	1979	1812–2145	302
3	S3F-MSA	1946	1784–2108	244
4	S2F-MSA	1941	1809–2072	105
5	ProSST-2048	1846	1708–1983	270
6	GVP-Mut	1790	1584–1995	84
7	Bridge-IF	1764	1542–1987	68
8	Knowledge-Design	1755	1526–1983	80
9	S3F	1700	1594–1807	116
10	TranceptEVE-L	1651	1561–1740	241

Remote sensing

#	Model	Score	95% CI	Votes
1	MSFMA-LGM	2325	2012–2639	94
2	LGRIN	2252	1895–2608	94
3	MobileNetV2 2.5x	2216	1949–2484	66
4	MobileViT-S	2012	1816–2208	66
5	MobileNetV2 2.0x	1959	1756–2163	66
6	EfficientFormerV2-S2	1911	1735–2088	66
7	LGNet	1855	1540–2171	78
8	RSMamba-H	1819	1649–1988	270
9	GhostNetV2 2.0x	1810	1605–2014	66
10	MobileViT-XS	1685	1558–1813	210

Agriculture

#	Model	Score	95% CI	Votes
1	SwinV2-B	1962	1782–2142	218
2	GPT-4o	1642	1536–1747	168
3	DenseNet-201	1567	1418–1715	242
4	Gemini-3-Flash	1541	1437–1644	255
5	Seed-1.6-Flash	1507	1406–1609	255
6	Gemini-2.5-Pro	1497	1420–1575	430
7	Gemini-2.5-Flash	1445	1347–1543	255
8	SCOLD	1436	1353–1520	270
9	Qwen3-VL-235B	1414	1317–1511	255
10	Qwen2.5-VL	1390	1315–1465	175

Chemistry / drug disc.

#	Model	Score	95% CI	Votes
1	STOUT v1.0	1599	1385–1813	60
2	InternS1-235B	1472	1293–1651	114
3	Gemini-3-Flash	1459	1288–1630	114
4	Uni-Mol	1426	1295–1556	108
5	DMFGAM	1386	1158–1613	84
6	ChemDFM-v2.0-14B	1287	1101–1472	114
7	GPT-5	1286	1124–1447	146
8	Kimi-K2	1221	1060–1382	179
9	MolXPT	1202	1026–1378	76
10	GEM	1191	1082–1300	195

Spatial transcriptomics

#	Model	Score	95% CI	Votes
1	STFlow	1791	1447–2135	56
2	GraphSAGE	1511	1340–1682	110
3	L-CAT	1480	1351–1609	110
4	GATv2Conv	1473	1338–1608	110
5	SGN	1418	1350–1485	348
6	TransformerConv	1358	1231–1485	110
7	STEM	1336	1237–1436	75
8	STAGE	1207	1098–1316	180
9	H-optimus-0	1199	1125–1273	305
10	M2OST	1194	1105–1284	75

H RELEASE AND A PUBLIC LEADERBOARD SITE

The artifact this method is for is not a table in a paper. We are building a leaderboard site that serves the fourteen boards as live rankings, and whose central feature is that every number is traceable: a score expands into the comparisons behind it, a comparison into its evaluation cell, and a cell into the named table, page and PDF that reported it, with the focal-model exclusion shown rather than assumed. It supports the head-to-head query the model actually answers — for two models, the battles they share and who reported them — and carries each board's collection window and identifiability status as first-class metadata, so a reader can see which rows are data-identified and over what period. Boards are refreshed as the window advances, which makes the stability measured in §6.3 a property the site has to keep rather than a one-off observation. The site is under construction at `https://bias-reduced-leaderboards.brl-leaderboards.workers.dev/`. It is not part of this submission and no result in the paper depends on it; every artifact it will serve is already in the release.