Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Capability, ranking dynamics, and inference pricing in the large language model ecosystem: evidence from linked public data

Peter Bago

peter.bago@uni-corvinus.hu

Corvinus University of Budapest

Systematic Review

Keywords: Large language models, Benchmark leaderboards, Scaling laws, Human preference, Price dispersion, Hedonic pricing

Posted Date: October 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11422935/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Capability, ranking dynamics, and inference pricing in the large language model ecosystem: evidence from linked public data

[First Author][1]*, [Second Author][2]

Peter Bago PhD, assistant professor, Corvinus University of Budapest, Hungary, 1093 Fővám tér 8. Institute of Entrepreneurship and Innovation, peter.bago@uni-corvinus.hu

---

Abstract. Public leaderboards and pricing catalogues offer complementary but hard-to-combine views of the large language model ecosystem. We analyse three linked public sources: an archived Open LLM Leaderboard v2 sample of 4,565 open-weight evaluation records with complete six-benchmark scores; an 89-snapshot human-preference panel of 279 models (June 2024–August 2025); and 2,685 priced chat endpoints (October 2026). Principal component analysis identifies a dominant common component (74% of benchmark variance), retained by parallel analysis and robust to family deduplication. Benchmark performance rises with parameter count (13.6 points per tenfold) and with submission date at constant size and type (10.5 points per year) under family-clustered inference. A documented change of the default rating on 20 May 2025 shifted all listed scores by about 18 points; within the stable period the preference frontier rose by about 187 points per year relative to a fixed reference set through 17 leaders with median spells of 12 days, top-20 slots concentrated among fewer organisations, and ratings did not drift within models. Among 201 models listed by three or more providers, the median maximum-to-minimum price ratio is 1.38 (1.27 under version-strict identifiers). A hedonic model adjusting for time since first listing attributes multipliers of about 3.1 to one standard deviation of arena score and 0.18 to open-weight status, robust to version-exact matching. Preference ratings and benchmark averages agree strongly for 69 matched models. Temporal mismatch, selective coverage, and related model families preclude causal interpretation.

Keywords: Large language models, Benchmark leaderboards, Scaling laws, Human preference, Price dispersion, Hedonic pricing

1. Introduction

The large language model (LLM) landscape has grown from a handful of frontier systems to thousands of publicly released models, hundreds of hosted application programming interfaces (APIs), and a continuous stream of human-preference comparisons. Since the transformer architecture [53] and few-shot learning at scale [7], capability has been tracked through benchmark suites [18, 30, 49], scaling laws [23, 27], and human-preference arenas [10, 58]. Most quantitative accounts of the market, however, rest on small samples of frontier models [36] and cannot speak to the long tail, to the dynamics of the ranking itself, or to how the price of inference is formed across competing providers.

This study assembles and links three large public sources: (i) an archived snapshot of the Open LLM Leaderboard v2 [16], which provides complete scores on six benchmarks for 4,565 open-weight evaluation records together with parameter counts, model type, licence, and submission date; (ii) an 89-snapshot panel of the LMArena text leaderboard covering 19 June 2024 to 6

August 2025, preserved through a public daily archive [5]; and (iii) the LiteLLM pricing catalogue [3], listing input and output prices, context windows, and feature flags for 2,685 chat endpoints across 82 provider categories. The sources are linked through normalised model names; every match is classified by whether the identifiers agree at version level, and the linkage tables are released with the paper. The analysis is observational; units, temporal alignment, and the dependence structure of the observations are documented explicitly because they determine which comparisons are meaningful.

Six questions organise the analysis:

RQ1. How are benchmark scores structured and distributed across a large sample of open-weight models—is there a dominant common component, and where in the score range do the benchmarks discriminate?

RQ2. How is capability associated with parameter count, with submission date at constant size and type, and with model type and licence?

RQ3. How has the human-preference frontier evolved relative to a stable reference set: growth, leadership turnover, the gaps between open-weight and proprietary and between US- and China-based developers, and the concentration of top-ranked slots among organisations?

RQ4. How do arena ratings behave as measurements: how does reported uncertainty scale with vote counts once repeated observations are accounted for, and do ratings change after a model enters once a documented change in the rating scale is accounted for?

RQ5. How are hosted inference prices formed: how dispersed are listed prices for the same model across providers, which features carry premiums, how does price scale with size, and what does a hedonic model attribute to capability and licensing?

RQ6. How closely do human-preference ratings and benchmark averages agree in the matched subset?

The contribution lies in the linkage and its documentation rather than in any single source: (i) a complete-score analysis of the benchmark structure of open-weight models with parallel analysis and family-adjusted checks; (ii) quantile and time-adjusted estimates of the size–capability association with dependence-robust uncertainty; (iii) a panel description of frontier dynamics measured against a fixed reference set within a documented period of rating-scale stability; (iv) a measurement of cross-provider price dispersion and feature premiums that tests the mechanical effect of provider count, in the tradition of the economics of information and hedonic pricing [1, 43, 50]; and (v) a cross-source agreement assessment. Three conclusions require at least two of the sources: the price–capability relationship (arena ratings and the price catalogue), the price–size relationship among open models (benchmark records and the price catalogue), and the agreement between preference ratings and benchmark averages.

2. Background

2.1 Benchmarks, leaderboards, and their incentives

The Open LLM Leaderboard v2 evaluated submitted open-weight models with a uniform harness on six benchmarks chosen for difficulty and contamination resistance: IFEval (instruction

following; [59]), BIG-Bench Hard (BBH; [52]), MATH Level 5 [19], GPQA [42], MuSR [48], and MMLU-Pro [55]. Each score is normalised between a random baseline and the maximum, with below-baseline results set to zero [16], which matters for the lower tail (Section 5.1). Because submission was open, the leaderboard also documents production practices—fine-tuning, weight merging in the spirit of model soups [57], mixture-of-experts architectures [15]—and the incentives of a public ranking: when a measure becomes a target it ceases to be a good measure [51].

2.2 The structure of LLM capabilities

Whether LLM capabilities form one dimension or several has been examined directly. Ilić and Gignac [25] report a strong general factor across a large set of models and benchmarks, with evidence of additional group factors; Burnell et al. [8] find a multidimensional factor structure on a different task set; and Ruan et al. [44] show that a low-dimensional capability space extracted from benchmark results predicts performance across model families, linking observational scaling to that space. Scaling laws [23, 27] and the foundation-model view of broad transfer [4] motivate the expectation of a dominant common component, while metric-dependent discontinuities [45] caution against over-reading any single benchmark. We add not a new factor model but a complete-score replication on a very large sample with explicit checks for family and size confounding.

2.3 Human-preference arenas as measurement instruments

Arena leaderboards convert pairwise human judgements into ratings via Elo [14], Bradley–Terry [6], or TrueSkill [20] models, with bootstrap confidence intervals [10]. In the published Bradley–Terry implementation the scale is fixed numerically by anchoring one reference model (Mixtral-8x7B-Instruct-v0.1 at 1114) whenever it is present [33]; numerical anchoring does not by itself guarantee comparability across dates, because the comparison set, the vote-processing rules, and the default rating variant can change [31], so absolute values from different dates must be compared with care (Section 5.4). A panel can ask whether reported uncertainty shrinks with votes near the inverse-square-root reference and whether ratings change after entry, provided that dependence from repeated listings is accounted for. Concentration of top-ranked slots among organisations is summarised with the Herfindahl–Hirschman index [21], here a measure of ranking concentration rather than market share.

2.4 Price formation for hosted inference

Hedonic methods decompose prices into implicit prices of characteristics [2, 17, 43], and quality-adjusted price declines have been documented for computing over long periods [39]. Hosted LLM inference adds a second classic question: when the same open-weight model is served by many providers, do listed prices converge, or do they remain dispersed as the economics of search and information predicts [1, 50]? Industry analyses of inference marketplaces have documented provider-level pricing and dispersion for individual models [11]; we add a catalogue-wide measurement. Algorithmic progress—capability gains at constant compute—has been estimated from training-compute records [22]; leaderboard data permit only a weaker quantity, the temporal association of capability with submission date at constant size and type.

3. Data

3.1 Sources, units, and sample construction

Table 1 summarises the three sources. The Open LLM Leaderboard v2 data are the archived snapshot dated 21 March 2025 of a daily CSV mirror [5]; it is the last snapshot the mirror holds and need not coincide with the final state of the original leaderboard. The unit is a leaderboard evaluation record identified by model name, commit hash, and numerical precision. The archive contains 4,576 records; we removed 1 record flagged by the leaderboard and 10 with missing or zero parameter counts, leaving 4,565 records for 4,486 distinct model names (79 names appear with two or more precision or commit variants, retained and treated as one family in the dependence-robust analyses). The uploading account (repository prefix) is not necessarily the developer; 10% of records come from accounts marked as official providers. The regression on submission date (Section 5.2) further excludes 5 records without a date and 9 records of rare types (multimodal, other), giving n = 4,551.

The LMArena panel comprises 89 dated snapshots of the text leaderboard (19 June 2024 – 6 August 2025) from the same mirror, with a median spacing of 4 days; each snapshot lists rank, arena score, 95% confidence interval, votes, organisation, and licence; the unit is a model-by-snapshot observation (15,224 observations on 279 distinct model names). Snapshots are irregularly spaced, so event timing is known only to within the interval. The 15,155 observations with a positive confidence-interval width and a positive vote count are used for the uncertainty analysis. The panel spans a documented change in the leaderboard's default rating view on 20 May 2025 (Section 5.4): 74 snapshots precede it and 15 follow, and cross-date comparisons of absolute scores are restricted accordingly. The LiteLLM catalogue is a community-maintained JSON file of provider price lists (retrieved 13 October 2026). Of 2,915 chat-mode entries, 2,761 carry both an input and an output price; we excluded 76 entries priced at zero on both sides (free tiers), leaving 2,685 priced endpoints from 82 provider categories, corresponding to about 76 economic entities after grouping regional and product variants of one vendor.

Table 1 Overview of the three data sources, their units of analysis, and their linkage

Attribute	Open LLM Leaderboard v2	LMArena panel	LiteLLM catalogue	Linkage
Unit	Evaluation record (model, commit, precision)	Model × snapshot	Endpoint (model × provider)	Normalised model name
Size	4,565 records; 4,486 names	279 models; 89 snapshots; 15,224 obs.	2,685 endpoints; 82 categories	Arena–price: 115; Arena–HF: 69; HF–price: 71
Period	Jun 2024 – Mar 2025	19 Jun 2024 – 6 Aug 2025	13 Oct 2026	See Section 3.3
Capability	6 normalised benchmarks, average	Arena score, 95% CI, votes	—	—
Covariates	Params, type, MoE, licence, base model, Hub likes, dates	Developer, licence, country (coded)	Prices, context, vision/reasoning/tool flags	—

3.2 Variables and coding

Open-ecosystem variables: six normalised benchmark scores (0–100), their average, parameters (billions, as recorded in the leaderboard metadata), model type (pretrained, continuously pretrained, fine-tuned, chat, merge), mixture-of-experts and merge flags, Hub licence (grouped into Apache-2.0, MIT, Llama, Gemma, CC-BY-NC, other, unspecified), Hub likes at the

snapshot date as a popularity measure, submission date, uploading account, and the declared base model, which defines model families: a record belongs to the family of its declared base model when that field names another repository; records whose base-model field is a placeholder ('Removed', 569 records) or the record itself form singleton families, giving 3,843 families, 159 of them with two or more records (881 records). Merges declare a single base model, so the other parents of a merge are not captured. The leaderboard's parameter field is derived from the safetensors metadata of the repository (element counts, with a factor-of-eight correction for GPTQ-quantised weights) [16]. We compared it with the size disclosed in the model name where a name states one and is not a mixture-of-experts name (3,392 records): 3.3% (112 records) differ by more than a factor of 1.5; in 28 of them the metadata count is close to half the named size, which is consistent with packed low-bit weights being counted by stored elements, but we could not verify repository contents, so these records are retained with the metadata value and all size-related estimates are repeated without them (Section 5.2). Arena variables: arena score, confidence-interval width (upper bound minus lower bound), votes, developer organisation, licence (coded open-weight when a non-proprietary licence is listed, proprietary otherwise; no record has a missing licence), and the developer's country coded from the organisation (United States, China, other), with the coding table released as supplementary material. Pricing variables: input and output price per million tokens, a blended price defined as their equal-weighted mean (a 1:1 token mix; 3:1 and 10:1 are reported as sensitivities), context window, and flags for reasoning, vision input, and function calling.

3.3 Linkage and temporal alignment

Names were normalised (lower case; provider prefixes, date stamps, and version suffixes removed) for linkage. Arena–price matches were accepted only when the normalised names were identical or differed solely by a recognised version suffix, and generic family names were excluded; this yields 115 matched models. Because the normalisation removes dates and version tokens, each match was then classified on the original identifiers: 72 pairs agree exactly (class A: the arena name, after separator normalisation, is itself one of the catalogue keys), 17 pair undated names on both sides (class B), and 26 pair a dated arena name with catalogue keys that do not carry that date (class C), where the catalogue alias may resolve to a different checkpoint; the cheapest listing is a class-A key for 61 models. The linkage table (original identifiers, catalogue keys, class) is released with the paper, and the hedonic estimates are repeated on class A alone. For each matched model the arena score is the value at the model's last listing in the panel; because the rating scale changed on 20 May 2025 (Section 5.4), we repeat the hedonic analysis with scores from the last snapshot before the change (13 May 2025) and from the final snapshot (6 August 2025). Time since first observed arena listing (in years, to the price retrieval date) is used as a covariate; for the 35 matched models already present in the first snapshot this quantity is left-censored, and results excluding them are reported. In the arena–leaderboard (69 models) and leaderboard–price (71 models) linkages, a model name maps to several evaluation records for 39 and 35 models respectively; we use the highest leaderboard average among those records and report the median as a sensitivity. The sources are not contemporaneous (benchmark records to March 2025, the panel to August 2025, prices from October 2026), so cross-source relationships describe how later list prices relate to earlier capability measurements, not a single market state.

4. Methods

Analyses used Python 3.13 with pandas [37], SciPy [54], statsmodels [46], scikit-learn [41], and Matplotlib [24]; α = .05, two-tailed. Procedures follow the estimand, the distribution of each variable, and the dependence structure of the observations. Benchmark scores were standardised to $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, and principal component analysis was applied to the Pearson correlation matrix R of the standardised scores [26],

$$\mathbf{R}\,\mathbf{v}_\ell = \lambda_\ell \mathbf{v}_\ell, \qquad \mathrm{EVR}_\ell = \frac{\lambda_\ell}{\sum_{m=1}^{p} \lambda_m} \tag{1}$$

reporting the full eigenvalue sequence, loadings defined as eigenvector elements scaled by the square root of the eigenvalue (variable–component correlations), Horn's parallel analysis against 500 random data sets, and the root-mean-square residual correlation of one- and two-factor maximum-likelihood solutions; the Spearman matrix was decomposed as a check. Internal consistency was summarised by Cronbach's α on the standardised scores [13],

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{j=1}^{k} \sigma_{Y_j}^2}{\sigma_X^2}\right] \tag{2}$$

Sensitivities repeat the decomposition on one random record per family and on residuals from regressions of each score on log parameters, and on log parameters, date, and type. Scaling was estimated by OLS and by quantile regression [28] of the leaderboard average on $\log_{10}$ parameters at $\tau \in 0.1, 0.5, 0.9, 0.99$, with percentile-bootstrap intervals from 2,000 resamples in which model families are resampled with replacement (seed 20261017; no fit failed; endpoints from the first 500 and 1,000 resamples agree to within 0.3 points),

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_i \rho_\tau\left(y_i - \mathbf{x}_i^\top \beta\right), \qquad \rho_\tau(u) = u\,(\tau - \mathbf{1}[u < 0]) \tag{3}$$

and the size- and type-adjusted temporal association by the linear model

$$\mathrm{Avg}_i = \beta_0 + \beta_1 \log_{10}(\mathrm{Params}_i) + \beta_2 t_i + \sum_k \gamma_k \mathrm{Type}_{ik} + \delta\,\mathrm{MoE}_i + \varepsilon_i \tag{4}$$

with standard errors computed three ways—HC3 heteroscedasticity-consistent [34], clustered by model family, and clustered by uploading account—with a replication on one record per family, on records excluding the parameter-count discrepancies, and on records that are themselves declared base models or pretrained checkpoints. The size × time interaction was estimated in a separate specification with both variables centred. Model types were compared with the Kruskal–Wallis test [29] as an omnibus test. In the arena panel, the frontier was measured in absolute score and relative to the mean score of a reference set of models present in at least 85 of the 89 snapshots (the mean is taken over the reference models present on each date; a model is the same reference only under an unchanged listing name), over the full panel and within the period before the rating-scale change; slopes of snapshot-level series carry Newey–West heteroscedasticity- and autocorrelation-consistent standard errors with four lags [38]. Common date effects were also estimated from a two-way model with model and date fixed effects on models with at least ten listings. Leadership spells were defined on the point-estimate rank at each snapshot (no ties occurred), with spells at the two ends of the panel censored. Concentration at date t was measured

with the Herfindahl–Hirschman index of organisations' shares of top-k ranked slots (k = 10, 20, 50) [21],

$$\mathrm{HHI}_t = \sum_{o=1}^{O} s_{ot}^2 \tag{5}$$

Rating uncertainty was examined with the log–log regression

$$\ln(\mathrm{CI\,width}_{it}) = a + b\,\ln(\mathrm{votes}_{it}) + \varepsilon_{it}, \qquad b_{\mathrm{ref}} = -\tfrac{1}{2} \tag{6}$$

with standard errors clustered by model, with and without date fixed effects, in a single-snapshot cross-section, and separately for snapshots before and after a documented change in the interval method; the hypothesis b = −1/2 was tested with the clustered standard error. Score change from a model's first to its last listing (models with ≥10 listings) was tested with the Wilcoxon signed-rank test [56], over the full panel and within the period of scale stability. Price dispersion for a base model m sold by providers p was summarised by the max/min dispersion ratio and by the standard deviation and interquartile range of log prices,

$$\mathrm{DR}_m = \frac{\max_p P_{mp}}{\min_p P_{mp}}, \qquad s_{\log P,m} = \sqrt{\frac{1}{n_m - 1}\sum_{p=1}^{n_m}\left(\ln P_{mp} - \overline{\ln P}_m\right)^2}, \qquad \mathrm{IQR}_{\log P,m} = Q_{0.75}(\ln P_{mp}) - Q_{0.25}(\ln P_{mp}) \tag{7}$$

Because the range of a sample grows mechanically with its size, the association between dispersion and provider count was re-examined on random subsamples of exactly three providers per model: each model's statistic is the median ratio over 200 draws, so that one value per model enters a single Spearman correlation with provider count. Dispersion was also recomputed under a stricter grouping in which identifiers differing in any date or version token are treated as distinct. Feature premiums were estimated by OLS of log price on feature flags and log context window with standard errors clustered by model and two-way by model and provider category [9], and the price–size relationship by

$$\ln P_e = \kappa_0 + \kappa_1 \log_{10}(\mathrm{Params}_e) + \varepsilon_e, \qquad \eta = \kappa_1 / \ln 10 \tag{8}$$

where η is the elasticity of price with respect to parameters. The hedonic model for arena-matched models was

$$\ln P_i = \gamma_0 + \gamma_1 z_i + \gamma_2\,\mathrm{Open}_i + \gamma_3\,\mathrm{CN}_i + \gamma_4\,\mathrm{Other}_i + \gamma_5\,\mathrm{Age}_i + \varepsilon_i \tag{9}$$

with z_i the arena score standardised on the full matched sample (so that within-segment coefficients share one scale) and Age_i the time since first observed listing; exponentiated coefficients are price multipliers under the log specification. Sensitivity analyses omit the five most influential observations by Cook's distance, add a quadratic age term, exclude left-censored models, use scores from the common snapshots named above, restrict the sample to version-exact matches, use the median instead of the minimum listed price, vary the input:output mix, drop one developer organisation at a time, and add a score × open-weight interaction. Group contrasts used the Mann–Whitney U test [35] with the rank-biserial correlation [12]; the Pareto-efficient set is defined by dominance in score and price; bivariate associations are Spearman correlations [47]. Cross-source agreement is reported with a percentile-bootstrap interval (2,000 resamples) and a permutation p-value (5,000 permutations).

4.1 Use of large language models

In accordance with the journal's policy, we document the use of a large language model (Claude, Anthropic; model family Claude 5, used between June and October 2026) during this study. The model assisted with data-extraction scripting, with drafting and refactoring of the analysis and figure code, and with language editing. All analytical decisions, specifications, and interpretations were made by the authors; every reported statistic was recomputed from the archived data by the authors' scripts, provided as supplementary material; the model was not used to generate, impute, or alter data and is not an author.

5. Results

5.1 Benchmark score distribution and common structure (RQ1)

The 4,565 evaluation records were uploaded by 715 distinct accounts, and the sample is dominated by derivative work: 39% of records are fine-tunes, 38% weight merges, 16% chat models, and 7% pretrained bases; 1.6% are mixture-of-experts. The ten most prolific accounts are individual uploaders (the most active submitted 197 records); the median record has 8.0 billion parameters (90th percentile 15.2B; maximum 141B), and 38% of records carry no Hub licence (32% Apache-2.0). Submissions totalled 4,560 over the period, peaking at 868 in January 2025.

Table 2 and Fig. 1 show low scores and limited separation near the lower end of the range, with no sign of a ceiling: no record scored 95 or more on any of the six normalised benchmarks (the maxima range from 29.4 on GPQA to 90.0 on IFEval). The normalisation sets below-chance results to zero, so the lower tail partly reflects the transformation [16]; in raw terms the median GPQA accuracy is 29% against a 25% chance level, with 5% of models at or below chance, 19% within two normalised points of the floor, and 42% within five; the median raw MATH Level 5 accuracy is 11%, with 3% of models at zero and 28% within five normalised points. Many records therefore achieve low but non-zero scores on the hardest benchmarks, and discriminating power in this sample is weakest at the bottom rather than at the top.

Table 2 Distribution of normalised benchmark scores (0–100) across the analysed records. The last two columns give the share of records scoring at most one point above the normalised floor and the share scoring at least 95

Benchmark	n	M (SD)	Mdn	P10–P90	Max	Skew	≤1 pt above floor	≥95
IFEval	4,565	45.6 (20.5)	45.2	18.4–74.8	90.0	0.05	0%	0%
BBH	4,565	27.6 (15.2)	29.5	4.9–49.2	76.7	−0.04	0%	0%
MATH Lvl 5	4,565	15.6 (14.6)	10.8	1.4–39.8	71.5	1.13	8%	0%
GPQA	4,565	6.7 (5.1)	5.9	1.0–14.3	29.4	0.89	9%	0%
MuSR	4,565	10.0 (5.8)	10.0	2.2–18.0	38.7	0.45	2%	0%
MMLU-Pro	4,565	25.4 (14.3)	27.2	2.4–46.6	70.0	−0.11	1%	0%

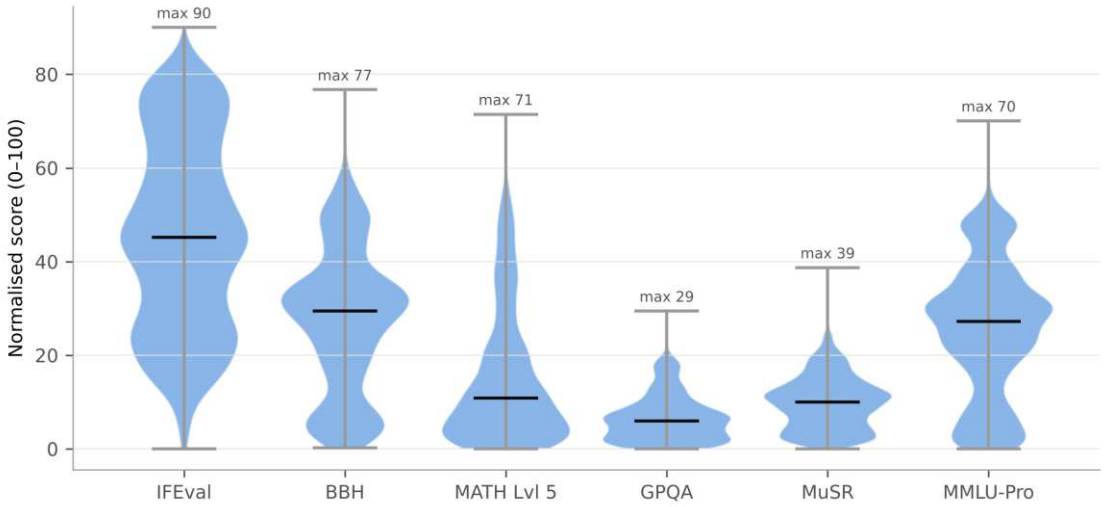


Fig. 1 Score distributions (violin plots with medians) for the six Open LLM Leaderboard v2 benchmarks (n = 4,565 records)

The correlation structure is dominated by a single component (Fig. 2). Pearson correlations among the six standardised scores range from 0.37 to 0.95 (Spearman mean 0.70); the eigenvalues of the Pearson matrix are 4.41, 0.74, 0.38, 0.27, 0.16, 0.04, so the first component explains 73.6% of variance (75.8% for the Spearman matrix), with loadings—variable–component correlations—between 0.69 (IFEval) and 0.96 (BBH, MMLU-Pro). Horn's parallel analysis retains exactly one component (95th-percentile random eigenvalues 1.07 and 1.04 against observed 4.41 and 0.74). A one-factor maximum-likelihood solution leaves a root-mean-square residual correlation of 0.058 (largest residual 0.12, IFEval–MuSR), reduced to 0.017 by a second factor that contrasts instruction following with multistep reasoning. Cronbach's α on the standardised scores is 0.925 (0.878 on the normalised scores).

A dominant component remained in the one-record-per-family analysis (n = 3,843 families; 73.3% explained) and after residualising the benchmark scores on log parameters (64.7%) and on log parameters, submission date, and type (63.1%); because residualised and re-standardised variables have a different denominator, these shares are not a decomposition of the original variance. The first component is larger among larger models: 64.8% (bootstrap 95% CI 62.7–66.9) at ≤4B parameters, 67.7% (66.6–68.7) at 4–15B, and 73.9% (71.7–76.5) above 15B. These results indicate a dominant common component among the six benchmark scores in the analysed sample, consistent with the general factor reported by Ilić and Gignac [25]; they do not establish that LLM capability is unidimensional in general, since the six benchmarks do not span all capabilities and the sample is a selection of submitted models.

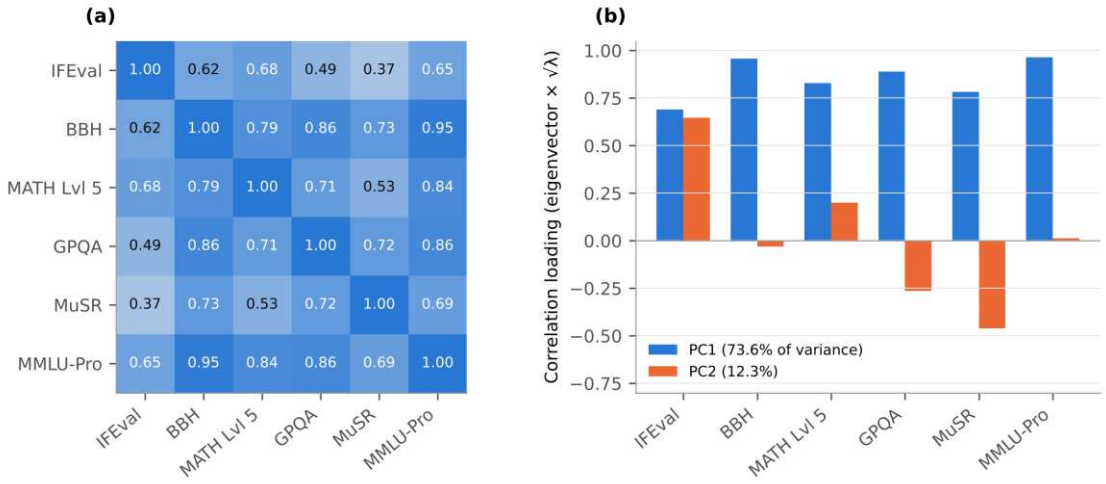


Fig. 2 (a) Spearman correlations among the six benchmark scores. (b) Correlation loadings (eigenvector elements scaled by the square root of the eigenvalue) of the first two components of the Pearson correlation matrix; first-component loadings range from 0.69 (IFEval) to 0.96 (BBH, MMLU-Pro)

5.2 Size, submission date, and model type (RQ2)

The leaderboard average is strongly associated with parameter count (Fig. 3): it rises by 13.7 points per tenfold increase in parameter count (SE 0.27; $R^2 = 0.40$; $\rho = 0.64$; $n = 4{,}565$). The association is steeper at the top of the conditional distribution: quantile-regression slopes are 7.8 points per tenfold at the 10th percentile (family-bootstrap 95% CI 5.8–9.9), 13.7 (12.9–14.2) at the median, 17.2 (16.5–20.5) at the 90th percentile, and 14.5 (12.2–17.4) at the 99th. The size-bin envelope (Table S6 of the Electronic Supplementary Material) shows diminishing but persistent gains in the maximum, from 22 (≤1B) to 52 (35–80B), with bin medians rising from 7 to 36; bin sizes differ by an order of magnitude, which raises expected maxima in dense bins.

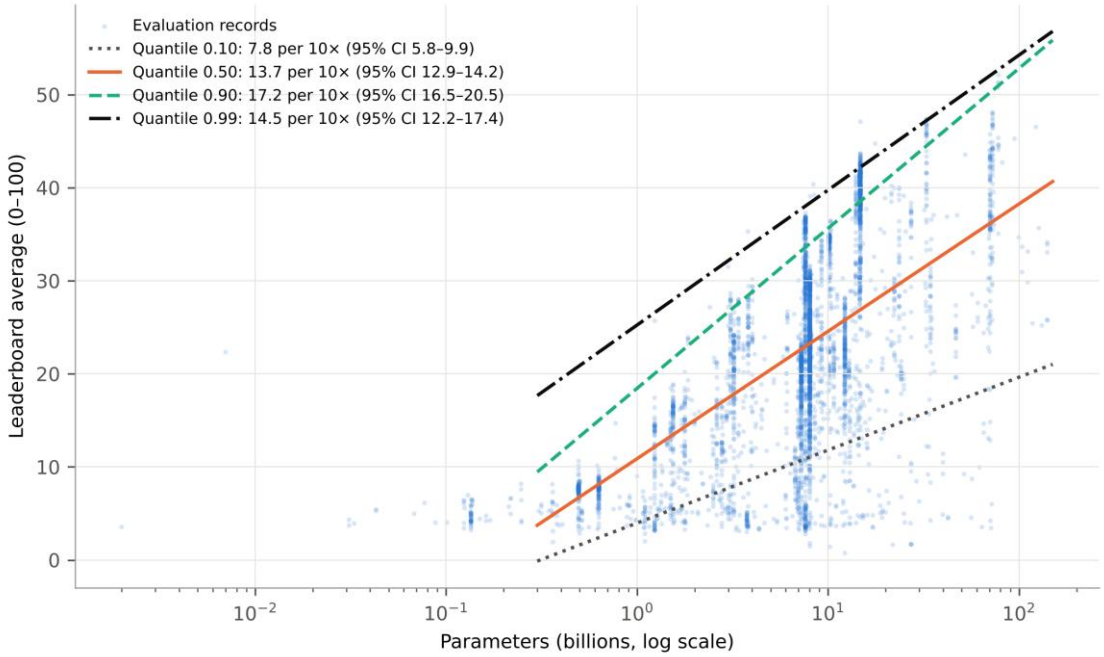


Fig. 3 Leaderboard average against parameter count (log scale) for 4,565 evaluation records, with quantile-regression lines at the 10th, 50th, 90th, and 99th percentiles; intervals are percentile bootstraps with model families resampled (2,000 resamples)

Conditional on size and type, later submissions score higher: 10.5 points per year of submission date (Table 3; Fig. 4). The estimate is unchanged under dependence-robust inference—HC3 SE 0.57; SE clustered by the 3,833 model families represented in the regression sample 0.67 (95% CI 9.2–11.8); SE clustered by the 713 uploading accounts 1.62 (CI 7.3–13.7)—and in a replication on one record per family (10.5 points per year, SE 0.61, n = 3,833). Excluding the 112 records whose metadata parameter count disagrees with the named size leaves the size slope at 13.7 and the date coefficient at 10.4 (10.4 when the 28 half-count records are instead set to the named size), and the first-component shares by size group move by less than one percentage point. Restricting the sample to records that are themselves declared base models or pretrained checkpoints (n = 1,318) halves the date coefficient (5.7, SE 0.99), so much of the association is carried by derivative submissions built on recent bases. We describe this as a size- and type-adjusted temporal association among submitted models: submission date is not release or training date, and the covariates do not hold training compute, data, or benchmark-oriented tuning constant, so the coefficient is not an estimate of algorithmic progress in the sense of compute-based studies [22]. With both variables centred, the size slope increases with submission date (interaction 11.4 points per $\log_{10}$ parameter per year, family-clustered 95% CI 7.5–15.4).

Relative to pretrained bases, chat models score 5.8 points higher, fine-tunes 3.4, and merges 6.0 (all $p < .001$; omnibus Kruskal–Wallis $H = 655$, $p < .001$), whereas mixture-of-experts models score 7.5 points lower at equal total parameters (family-clustered SE 0.85), consistent with capability tracking active rather than total parameters [15]. Merging combines existing checkpoints without additional training [57], yet merges match or exceed fine-tunes at equal size, and merged or fine-tuned derivatives top every size bin (Table S6)—a pattern consistent with

several explanations, including benchmark-oriented optimisation of the kind anticipated by Goodhart-type arguments [51], which the data do not identify.

Table 3 OLS model of the leaderboard average (n = 4,551; R^2 = 0.494; HC3 standard errors). Family- and uploader-clustered standard errors for the date and MoE terms are given in the text; the size × date interaction is estimated in a separate centred specification

Term	β	SE (HC3)	p	Interpretation
Intercept	1.26	0.55	.021	—
$\log_{10}$ parameters	13.60	0.31	<.001	points per tenfold parameters
Years since June 2024	10.51	0.57	<.001	points per year at constant size and type
Chat (vs. pretrained)	5.76	0.53	<.001	type contrast
Fine-tuned (vs. pretrained)	3.41	0.49	<.001	type contrast
Merge (vs. pretrained)	6.01	0.52	<.001	type contrast
Mixture-of-experts	−7.54	0.81	<.001	at equal total parameters

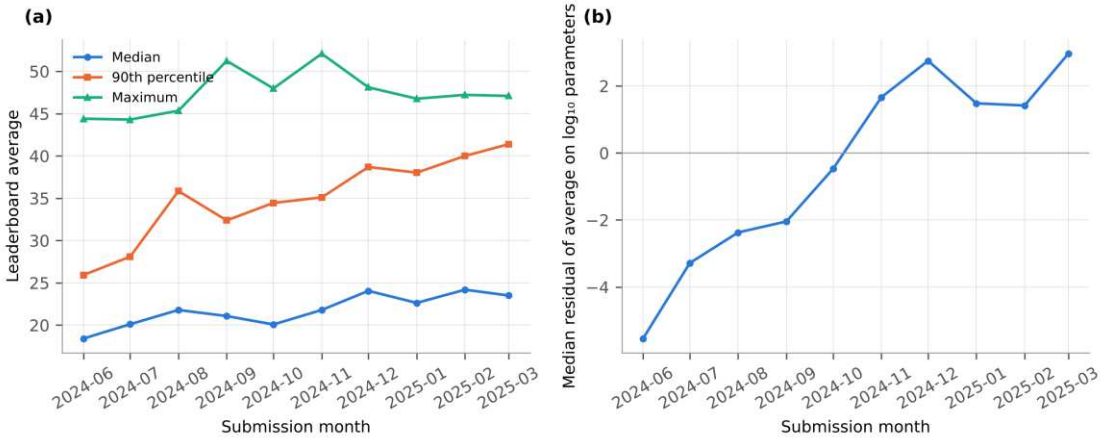


Fig. 4 (a) Median, 90th percentile, and maximum leaderboard average by submission month (months with at least 20 records). (b) Median residual by submission month from an OLS regression of the leaderboard average on $\log_{10}$ parameters over all dated records

Hub likes are weakly coupled to measured capability (ρ = 0.13 with the leaderboard average, −0.01 after adjusting for size); the median average of the 100 most-liked models sits at the 36th percentile. Licence differences at equal size are small: relative to Apache-2.0 records, CC-BY-NC (−2.5) and Llama-licensed (−2.2) records score lower (both $p < .001$), with no significant difference detected for MIT- or Gemma-licensed records.

5.3 Frontier dynamics, gaps, and ranking concentration in the arena panel (RQ3)

Over 89 snapshots the leaderboard grew from 109 to 260 listed models (54% open-weight at the end); the final snapshot lists 161 models from US-based developers, 55 from China-based, and 19 from France-based developers. The highest listed score rose from 1287 to 1472 (Fig. 5a), a regression slope of 189 points per year (Newey–West SE 5.1; an endpoint-to-endpoint rate would be 164). Because absolute scores are comparable across dates only while the rating scale is stable (Section 5.4), we also measure the frontier relative to a reference set of 92 models listed in at least 85 of the 89 snapshots (86–93 present per date, median 92). Over the 74 snapshots before

the change of 20 May 2025, the reference mean was flat (1076–1078 after the first snapshot) and the frontier rose by 188 points per year in listed scores (SE 7.1) and by 187 relative to the reference mean (SE 7.3; Fig. 5b). Over the full panel the relative slope is 179 points per year (SE 6.3), because the reference mean itself shifted from 1078 to 1103 after the change (176–179 under alternative reference definitions). A two-way model with model and date fixed effects on models with at least ten listings gives common date effects within ±0.5 points before the change and 20 points at the end of the panel; a linear slope through them (9.2 points per year, 10 over the panel) understates this step-shaped series. The reference adjustment removes the common level shift but does not separate capability change from all changes in evaluation composition or rating calibration.

First place changed hands 19 times across 20 spells held by 17 distinct listed names (9 after collapsing dated or 'experimental' variants of the same product line), with no ties at the point estimate (the leaders are listed in Table S5 of the Electronic Supplementary Material). The median spell lasted 12 days (10 days excluding the censored first and last spells; 5 spells were observed at a single snapshot), snapshot spacing (median 4 days) bounds the precision of these durations, and first place by point estimate does not imply a statistically distinguishable lead. Leadership alternated among OpenAI (through late 2024), Google (most of 2025), and briefly xAI, while the gap between the first- and tenth-ranked model stayed roughly constant (57 points at the start, 60 at the end).

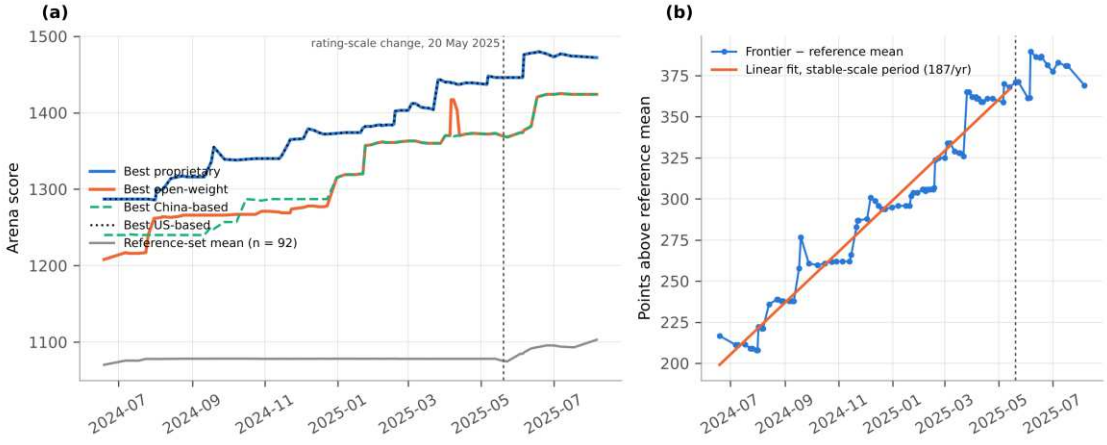


Fig. 5 (a) Best arena score by licensing regime and by developer country over time, with the mean of the 92-model reference set (grey) and the rating-scale change of 20 May 2025 (dotted line); from January 2025 the best open-weight and best China-based lines coincide on all but three snapshot dates. (b) Frontier minus reference mean, with a linear fit over the stable-scale period

The best open-weight model trailed the best proprietary model by 79 points in June 2024, by 21 points at the narrowest, and by 48 points at the end (mean 59); the fitted linear trend is −0.8 points per year (Newey–West 95% CI −23 to 21; −10.5, CI −40 to 19, within the stable-scale period), so no linear trend was detected, and the intervals are wide enough to admit both narrowing and widening. The best China-based developer's model trailed the best US-based one by 47 points at the start and 48 at the end (minimum 21; trend −2.5 per year, CI −23 to 18). Gaps and ranks are differences within a snapshot and are unaffected by the common level shift. The two gaps coincided when the leading models of the two groups were identical: from January 2025

the best open-weight and the best China-based model were the same system (the DeepSeek line) on 47 of 50 snapshot dates, the exceptions being three April 2025 listings of 'Llama-4-Maverick-03-26-Experimental' (Meta; open-weight, US-based), which reached 1417 and then disappeared. Because many lower-scoring open models enter continuously, the gap between the median open and the median proprietary model widened (100 → 125 points), a compositional change not decomposed here.

Concentration of top-ranked slots increased (Fig. 6): the Herfindahl–Hirschman index of organisations' shares of top-20 slots rose from 0.16 to 0.25, and the number of organisations in the top 20 fell from 10 to 6; the pattern holds for the top 50 (0.11 → 0.18; 14 → 11 organisations) but not for the top 10, where the index fell (0.34 → 0.24) as a fifth organisation entered. The index of cumulative votes by organisation was stable (0.11 → 0.11). These are measures of ranking concentration, not of market shares in revenue or usage. The share of top-20 slots held by China-based developers moved from 20% to 15% and the US-based share from 75% to 85%, while the open-weight share remained 20%. In the final snapshot, open-weight models score a median 1163 against 1288 for proprietary models (rank-biserial r = 0.55, p < .001).

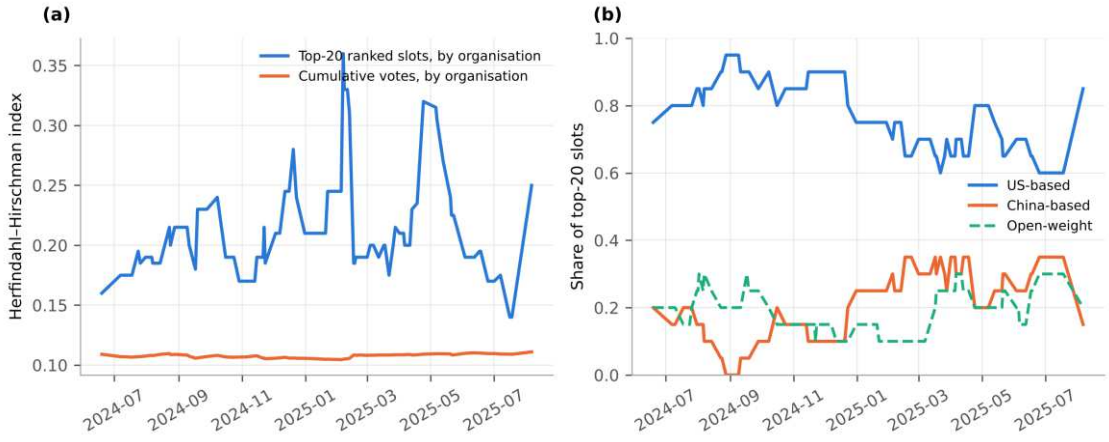


Fig. 6 (a) Herfindahl–Hirschman indices of top-20 ranked slots and of cumulative votes by organisation. (b) Shares of top-20 slots held by US-based, China-based, and open-weight models

5.4 Arena ratings as measurements (RQ4)

Confidence-interval width declines with vote count with an estimated log–log slope of −0.432 (Fig. 7a). A naive regression on the 15,155 model-date observations gives a standard error of 0.001, but observations of the same model are strongly dependent; clustering by model raises the standard error to 0.006 (95% CI −0.443 to −0.421; −0.433 with date fixed effects). The operator replaced bootstrap intervals with a closed-form method and introduced vote reweighting on 23 July 2025 [31]; the 88 snapshots before that date give −0.432 (clustered SE 0.006), the last cross-section before it (18 July 2025, n = 218) −0.473 (SE 0.008), and the final snapshot (n = 260) −0.427 (SE 0.010). Under all specifications the slope is shallower than the inverse-square-root reference value of −0.5 (p < .001 with the clustered standard error). The reference value is an idealisation for independent observations; a Bradley–Terry interval depends on the pairing network and win proportions, and non-uniform pairing is one plausible reason for the shallower slope, offered as a hypothesis rather than a finding.

The anchor model of the published rating implementation, Mixtral-8x7B-Instruct-v0.1, is listed at exactly 1114 in every one of the 74 snapshots from 19 June 2024 to 13 May 2025, as the anchoring rule implies [33], and at 1111–1138 in the 15 snapshots from 20 May 2025, after the operator announced that the style-controlled rating would become the default leaderboard view [32]. Across this change the scores of the 235 models listed on both sides shifted by a median of +18 points (interquartile range 16–23) between the last snapshot before it and the end of the panel, which is the level shift visible in the reference mean and in the date effects of Section 5.3. Within the stable-scale period, ratings did not drift: among the 228 models with at least ten listings before 20 May 2025, the median change from first to last listing is 0 points (35% rose, 29% fell; Wilcoxon p = .586), whereas over the full panel the same statistic is +19 points with 95% of models rising (Fig. 7b)—a difference produced by the rating-scale change, not by within-model drift. Longitudinal uses of archived arena scores should therefore treat 20 May 2025 as a break in the scale, as we do, or rely on rank- and gap-based measures, which are unaffected by a common shift.

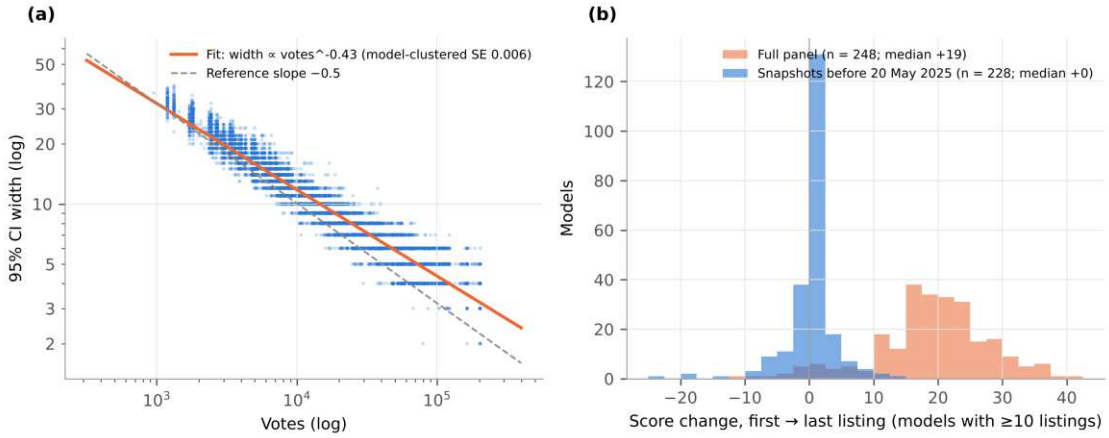


Fig. 7 (a) Confidence-interval width against votes (log–log) with the fitted power law (model-clustered standard error) and the −0.5 reference slope. (b) Score change from first to last listing for models with at least ten listings: full panel versus snapshots before the rating-scale change of 20 May 2025

5.5 Price formation for hosted inference (RQ5)

Across 2,685 priced endpoints the median input and output prices are $0.60 and $1.68 per million tokens (median blended price $1.15; 10th–90th percentile $0.16–$10.46; median output/input ratio 4.0); 47% are flagged reasoning-capable and 41% accept images. Feature premiums are large (Fig. 8b; n = 2,579 endpoints with a recorded context window): holding the others constant, vision input carries a +138% premium and each tenfold increase in context window +55% (both $p < .001$ with standard errors clustered two-way by model and provider category), reasoning capability +43% (two-way clustered p = .042; $p < .001$ under HC3), whereas the function-calling coefficient is not statistically significant (+16%, two-way clustered p = .530; R^2 of log price = 0.22). Hosted prices also follow model size: among 884 endpoints whose normalised key discloses a parameter count (the largest number followed by ’b’, so that mixture-of-experts names such as 8x7b contribute the per-expert size; 24 such endpoints), price rises by +229% per tenfold parameters (elasticity $\eta = 0.52$; model-clustered SE of the $\log_{10}$ slope 0.076; $R^2 = 0.36$; Fig. 8a). Two endpoints of the Jais models (Azure AI, watsonx) are listed at blended prices of

$6,455 and $1,250 per million tokens, two to three orders of magnitude above the next-highest endpoint in this sample ($10.68) and far above the highest other price in the catalogue ($112.50); we could not verify these catalogue entries, and excluding them leaves the slope at 1.20 and raises R^2 to 0.42 (excluding mixture-of-experts names instead gives 1.22). At the level of the cheapest listing per model the elasticity is 0.57 ($R^2 = 0.47$; n = 112). In the matched open-weight sample of 71 models, log price is associated with log parameters ($\beta = 1.44$, $p < .001$) whereas the benchmark-average coefficient is not distinguishable from zero ($\beta = -0.0033$ per point, 95% CI −0.026 to 0.019, i.e., −25% to +24% per standard deviation; $R^2 = 0.43$), so a moderate quality effect is not excluded.

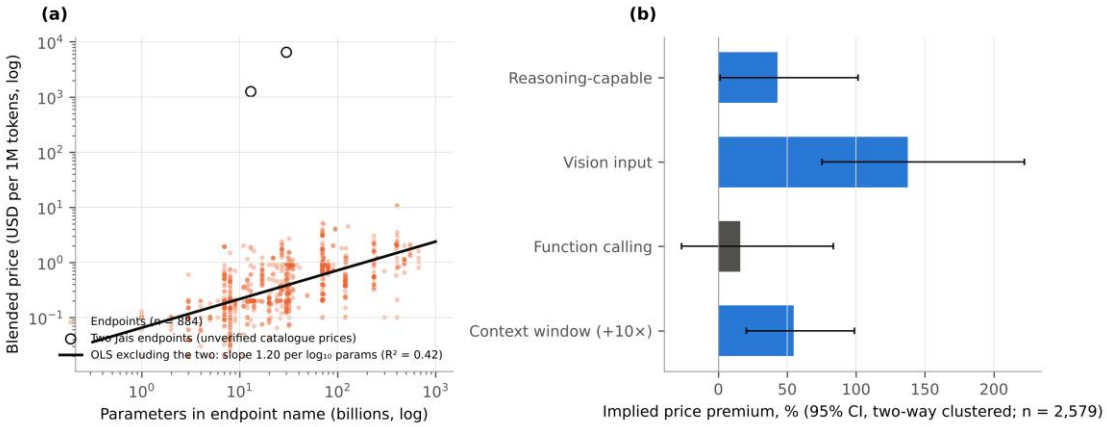


Fig. 8 (a) Blended price against the parameter count disclosed in the endpoint name (n = 884), with the two unverified Jais endpoints marked and the OLS fit excluding them. (b) Implied price premiums for feature flags and context window with 95% intervals from standard errors clustered two-way by model and provider category (grey = not significant at the 5% level)

Listed prices for the same catalogue model differ across providers. Among 201 models offered by three or more provider categories (178 after grouping categories into economic entities), the median ratio of the highest to the lowest blended price is 1.38 (1.56 at entity level), the upper quartile 4.0 and the 90th percentile 8.0; 39% of models show at least a twofold spread, the median standard deviation of log prices is 0.14, and the median interquartile range of log prices 0.10 (Fig. 9). The median ratio implies that switching from the most to the least expensive listing saves 27% of the higher price (1 − 1/1.375); this is an upper bound on the saving available to a buyer, whose expected saving depends on the distribution of starting providers. Under the stricter grouping that treats identifiers differing in any date or version token as distinct (192 models), the median ratio is 1.27, 36% of models exceed twofold, and the median log-price SD is 0.12; the lenient grouping merges 159 of 1,311 identifier groups that contain more than one version-level identifier, so part of the measured dispersion may reflect version differences. The ratio is stable across input:output mixes (1.38–1.42 from 1:1 to 10:1). Dispersion increases with the number of providers ($\rho = 0.39$ for the max/min ratio; 0.30 for the log-price SD, $p < .001$): the most widely hosted model, gpt-oss-120b, is listed by 19 providers at $0.10 to $0.80. Part of this association is mechanical, because the range of a sample grows with its size; on random subsamples of exactly three providers per model the median ratio falls from 3.98 to 1.86 for models with six or more providers, yet the correlation between provider count and the per-model median subsample ratio remains positive at fixed sample size ($\rho = 0.29$, bootstrap 95% CI 0.16 to 0.41, $p < .001$; Fig. 9b). The cross-section cannot show whether dispersion changed as providers entered, and identical

weights need not be identical services (latency, throughput, quantisation, region, and caching can differ), so these figures describe dispersion of listed prices rather than a test of search-cost equilibrium.

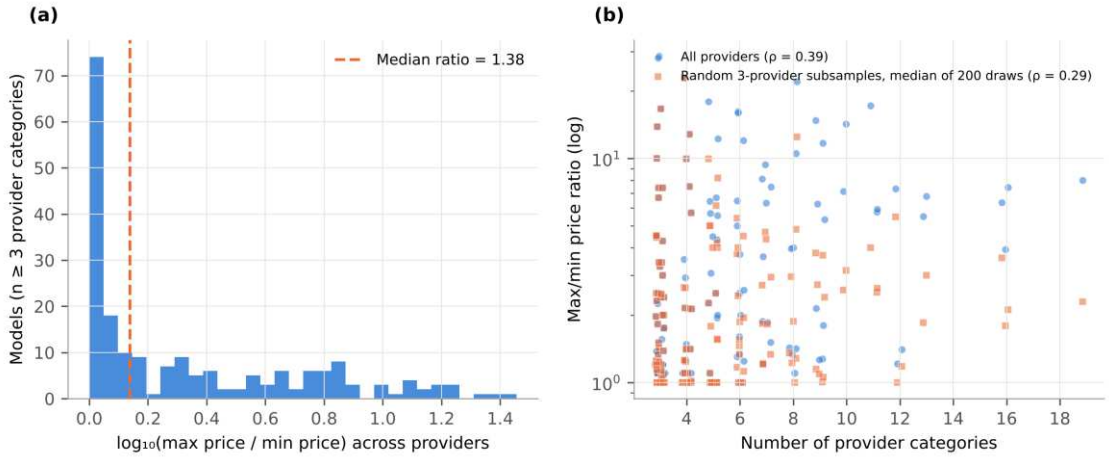


Fig. 9 Cross-provider price dispersion for the 201 models with three or more provider categories. (a) Distribution of the max/min price ratio. (b) Ratio against number of provider categories: all providers (circles) and the per-model median over 200 random three-provider subsamples (squares), with the respective Spearman correlations

For the 115 arena-matched models (Fig. 10; Table 4), arena score and lowest listed price are only weakly related unconditionally ($\rho = 0.32$), and without a time term the score coefficient is not statistically significant ($p = .175$) while open-weight status is associated with an 88% lower listed price (multiplier 0.12). Adding time since first observed arena listing raises the explained variance of log listed price to 54%: one standard deviation of arena score is associated with a price multiplier of 3.13 ($\gamma_1 = 1.14$, 95% CI 0.69 to 1.59, $p < .001$), open-weight status with a multiplier of 0.18 (−82%; $p < .001$), and each additional year since first listing with a multiplier of 16.88 ($\gamma_5 = 2.83$, $p < .001$). Score and time since listing are strongly negatively correlated ($\rho = -0.81$), so the time coefficient is a conditional cohort difference—earlier-listed models have higher current list prices at a given score—rather than evidence that individual prices persisted; a quadratic term indicates a convex association (age −8.9, age² 3.3, $p = .006$). The China-based term is not statistically distinguishable from zero in this specification ($p = .741$); 17 of the 24 China-based models are open-weight, so the two indicators are partly collinear even though variance inflation factors [40] remain at or below 3.63. The coefficients are robust (Table 5): omitting the five most influential observations (Cook's distance), excluding the 35 left-censored models, taking scores from the last snapshot before the rating-scale change or from the final snapshot, restricting the sample to the 72 version-exact matches (or the 61 whose cheapest listing is a version-exact key), using the median instead of the minimum listed price, using 3:1 or 10:1 input:output mixes, and dropping any one of 18 developer organisations leave the score coefficient between 0.97 and 1.67 and the open-weight coefficient between −2.33 and −1.03. Within segments, using the common standardisation, the score coefficient is 1.29 (95% CI 0.40 to 2.19; multiplier 3.6) among proprietary models ($n = 50$) and 0.84 (0.35 to 1.33; multiplier 2.3) among open-weight models ($n = 65$); a score × open-weight interaction in the pooled model is not statistically significant ($\gamma = -0.35$, SE 0.28, $p = .206$), so the two segment estimates are not distinguishable. The Pareto frontier of score against lowest listed price contains 8 models, 6 of

them open-weight; because the minimum listed price falls with the number of providers, this frontier describes the cheapest available listing rather than a licence effect.

Table 4 Hedonic regression of ln(lowest blended list price) for arena-matched models (n = 115; $R^2 = 0.540$; VIF ≤ 3.63). Multipliers are price ratios under the log specification

Term	γ	SE (HC3)	p	95% CI	Multiplier exp(γ)
Intercept	−4.908	1.089	<.001	[−7.04, −2.77]	—
Arena score (+1 SD)	1.141	0.230	<.001	[0.69, 1.59]	3.13
Open-weight	−1.699	0.287	<.001	[−2.26, −1.14]	0.18
China-based developer	0.079	0.238	.741	[−0.39, 0.55]	1.08
Other region	0.505	0.365	.166	[−0.21, 1.22]	1.66
Years since first listing	2.826	0.536	<.001	[1.78, 3.88]	16.88

Table 5 Sensitivity of the hedonic coefficients to influential observations, censoring, score timing, version-level linkage class, price definition, token mix, and developer organisation (HC3 standard errors). All score and open-weight coefficients remain significant at $p < .01$

Specification	n	R^2	γ score	γ open-weight	γ years since listing
Baseline (Table 4)	115	0.54	1.14	−1.70	2.83
Without five most influential (Cook's D)	110	0.58	1.08	−1.65	2.87
Excluding left-censored models	80	0.57	1.67	−1.44	2.74
Scores from last pre-change snapshot (13 May 2025)	84	0.60	1.07	−2.33	3.44
Scores from final snapshot (6 Aug 2025)	105	0.55	1.07	−1.77	2.73
Version-exact matches only (class A)	72	0.57	1.18	−1.94	3.05
Classes A and B	89	0.50	1.00	−1.85	2.69
Median instead of minimum price	115	0.49	1.25	−1.03	2.84
3:1 input:output mix	115	0.51	1.12	−1.55	2.89
10:1 input:output mix	115	0.47	1.12	−1.38	3.00
Leave-one-organisation-out (range)	—	—	0.97 to 1.37	−1.91 to −1.34	2.36 to 3.31

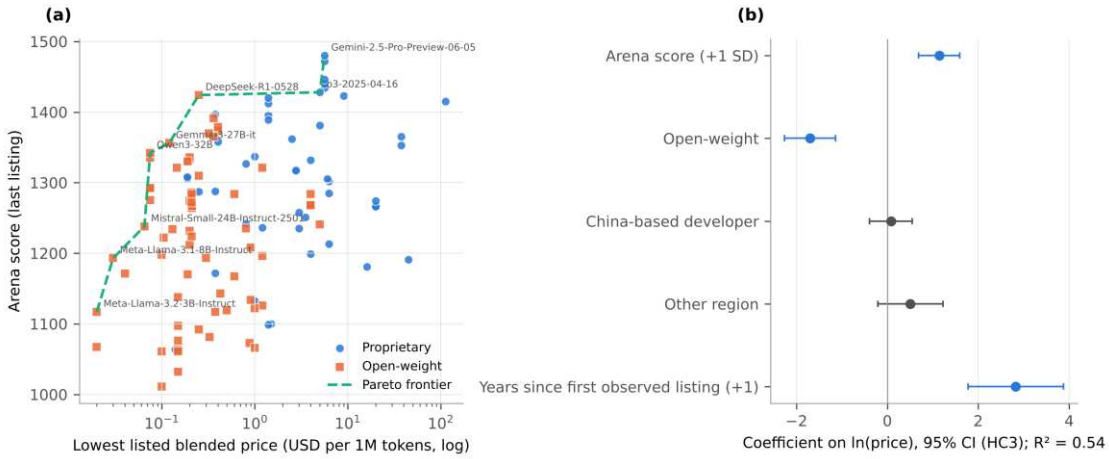


Fig. 10 (a) Arena score at last listing against lowest listed price with the Pareto frontier. (b) Hedonic coefficients with 95% HC3 intervals (grey = not significant)

5.6 Cross-source agreement (RQ6)

For the 69 models present in both the arena panel and the Open LLM Leaderboard (1.5% of the leaderboard records), the arena score correlates $\rho = 0.85$ with the six-benchmark average (bootstrap 95% CI 0.74 to 0.92; permutation $p < .001$; Fig. 11). The correlation is 0.77 when the median rather than the maximum leaderboard average is used for the 39 names with several records, and 0.85 with each model's peak arena score. This is evidence of convergent ranking agreement within the matched subset, which is composed of widely used models and may share size, family, and cohort influences; it does not establish a common scale, and pooling the two sources would require calibration and external validation that we do not attempt.

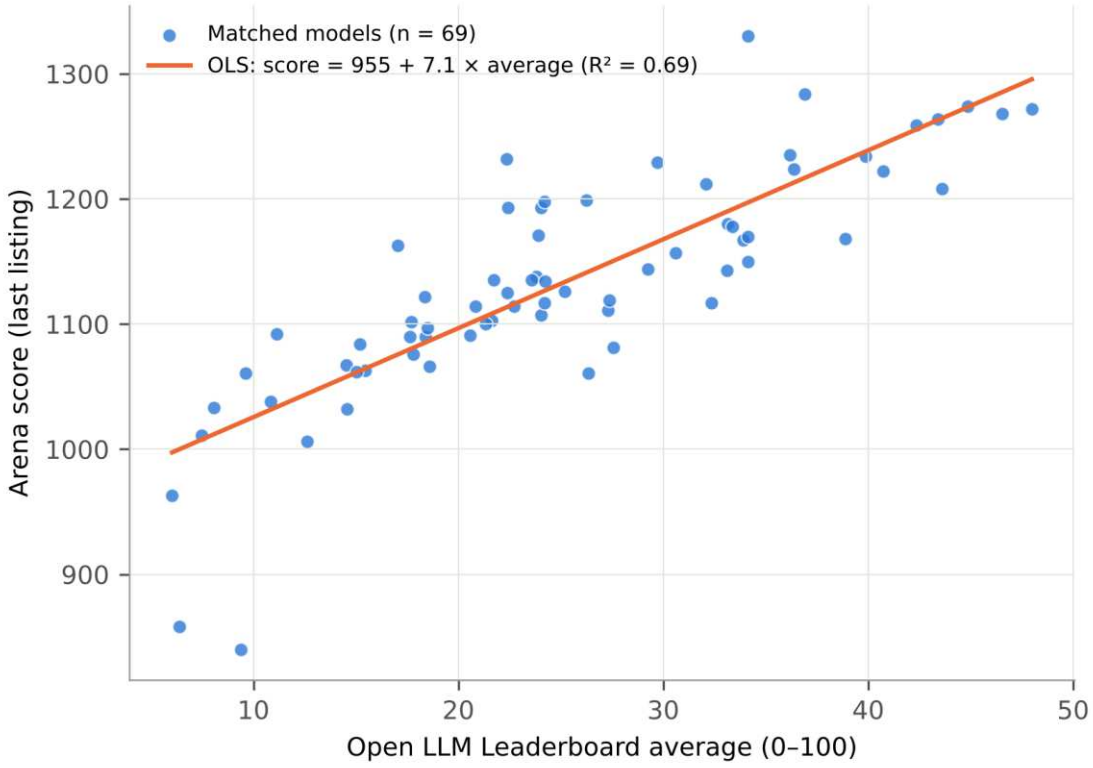


Fig. 11 Arena score at last listing against the Open LLM Leaderboard average for the 69 models present in both sources, with the OLS line; the Spearman correlation reported in the text is a separate rank-based statistic

6. Discussion

Six observational findings follow from analysing linked public records rather than a frontier sample. First, the six benchmark scores of the analysed open-weight records exhibit a dominant common component—74% of variance, retained alone by parallel analysis, $\alpha = 0.92$—that remains in one-record-per-family and size-adjusted analyses and strengthens with model size, consistent with the general factor of Ilić and Gignac [25] and the low-dimensional capability space of Ruan et al. [44], while other task sets yield multidimensional structures [8] and our six benchmarks do not span all capabilities. Second, benchmarks that saturate at the frontier show low scores and limited separation at the lower end of the ecosystem, partly shaped by zero-clipping normalisation; a single leaderboard cannot serve both the frontier and the long tail.

Third, size and submission date are both strongly associated with capability—13.7 points per tenfold parameters, steeper at the upper quantiles, and 10.5 points per year at constant size and type—with clustered inference and the exclusion of parameter-count discrepancies leaving the estimates intact, although the date coefficient halves among records that are themselves base models. The temporal association is a property of submitted records, not an estimate of algorithmic progress [22]; the dominance of merges and fine-tunes in every size bin and the weak coupling of Hub likes to size-adjusted capability are consistent with benchmark-oriented optimisation among other explanations.

Fourth, the human-preference frontier rose by about 187 points per year relative to a fixed reference set while the rating scale was stable (179 over the full panel; 189 in listed scores), through 17 listed names (9 product lines) with median spells under two weeks, while the number of organisations in the top 20 fell from 10 to 6—rising ranking concentration, not market concentration. The open–proprietary and China–US gaps coincided while the leading open and leading China-based model were the same system, and neither showed a statistically significant trend. Fifth, interval widths shrink with votes more slowly than the inverse-square-root reference once repeated listings are accounted for; ratings do not drift within models while the scale is stable, but the change of the default rating on 20 May 2025 shifted all listed scores by about 18 points, so cross-date comparisons of archived scores must respect that break or use relative measures. Sixth, associations between listed prices, model characteristics, and measured capability differed across samples and specifications. Prices for the same catalogue model differ by a median factor of 1.38 across providers (1.27 under version-strict identifiers), by more than twofold for 39% of multi-provider models, and more for widely hosted models even at fixed sample size—consistent with, but not a test of, search-cost explanations [1, 50]. Hosted prices scale with parameters ($\eta \approx 0.52$), while in the matched open-weight sample the benchmark coefficient is imprecisely estimated. In the pooled arena-matched sample a hedonic model attributes a multiplier of about 3.1 to a standard deviation of arena score (3.6 among proprietary and 2.3 among open-weight models, not significantly different), 0.18 to open-weight status, and a large cohort premium for earlier-listed models that is consistent with list-price persistence but not direct evidence of it. For procurement, a search across hosts of the same catalogue model yields a median saving of up to 27% relative to the most expensive listing; for benchmark design, tiered suites and floor-aware normalisation; for competition analysis, frequent leadership turnover coexisting with rising ranking concentration.

7. Limitations

(1) Selection and units: the Open LLM Leaderboard records models submitted for evaluation, over-representing derivative work; records are evaluation configurations rather than unique models (79 names appear in several variants), metadata parameter counts disagree with name-disclosed sizes in 3.3% of checkable cases and could not be verified against repository contents, the uploading account is not the developer, and the leaderboard was retired in 2025. Family-clustered errors and one-per-family replications address dependence, but the declared base model is an imperfect family identifier (a placeholder for 569 records; merges declare one parent; common ancestry beyond the declared base is not captured). (2) The arena panel mirrors published tables with irregular spacing; models are occasionally removed or renamed, and country coding from organisation names is approximate. The panel spans a documented change of the default rating on 20 May 2025; we report estimates within the stable period, reference-

anchored and fixed-effects measures, and rank-based statistics, but the archived tables allow neither calibration against a common vote set nor identification of the mechanism of the shift. (3) The sources are not contemporaneous (benchmarks to March 2025, arena to August 2025, list prices in October 2026 from a community catalogue); time since first listing partly absorbs the mismatch but is left-censored for 35 models and observed at a single price, so cohort and persistence effects cannot be separated. List prices may differ from transaction prices; provider categories are not independent economic actors; 26 of the 115 arena–price pairs link a dated arena name to an undated catalogue alias that may resolve to a different checkpoint, so results on version-exact matches are reported. (4) Parameter counts parsed from endpoint names omit active-parameter and undisclosed sizes, and two catalogue prices could not be verified. (5) All relationships are observational: the temporal association conflates methodological change with changes in who submits, type contrasts reflect what submitters chose to build, and the dispersion analysis is a single cross-section. Replication with raw evaluation files, transaction-level and historical prices, and a successor open leaderboard would strengthen all three strands.

8. Conclusion

Linking public benchmark, preference, and pricing records yields complementary observational patterns: a dominant common component among the benchmark scores of 4,565 open-weight records, with low scores rather than ceilings at the hard end; strong associations of capability with size and with submission date; a human-preference frontier rising by about 187 points per year relative to a fixed reference while the rating scale was stable, with short leadership spells and rising ranking concentration; an open–proprietary gap that coincided with the China–US gap and showed no detectable trend; arena ratings whose uncertainty shrinks with votes more slowly than the independent-sampling reference and whose scale shifted by about 18 points at a documented change in May 2025 without drifting otherwise; and listed inference prices that vary substantially across providers of the same catalogue model and are associated with size, licensing, and listing cohort as well as with preference ratings. These results describe selected public records rather than the population of deployed systems. Progress requires version-aware identifiers, temporally aligned measurements, documented rating-scale changes, and uncertainty estimates that account for related models and repeated observations; maintaining benchmark, preference, and price data in reproducible, linkable form would improve the evidence available for evaluation and procurement decisions.

Statements and Declarations

Funding

No funding was received for conducting this study.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable: the study uses publicly available model-level data; the human-preference votes underlying the arena ratings were collected and published by the arena operators.

Data availability

All three sources are public: the Open LLM Leaderboard v2 and LMArena snapshots through the daily CSV mirror [5] (archived files are identified by snapshot date and commit hash in the supplementary manifest), and the LiteLLM pricing catalogue through its GitHub repository [3] (retrieved 13 October 2026; the retrieved file is archived with the supplementary material). The processed data sets, the linkage tables with original identifiers, catalogue keys and version-level match classes, the country-coding table, the snapshot manifest, and the complete analysis and figure code (with package versions and random seeds) are provided as Electronic Supplementary Material and will be deposited in a public repository with a persistent identifier upon acceptance.

Declaration

During the preparation of this manuscript, the authors used ChatGPT (OpenAI; model: Astra Max) to assist with translation into English. The AI-generated translations were reviewed and, where necessary, revised by the authors to preserve the intended scientific meaning, ensure the accuracy of technical terminology, and maintain clarity and consistency throughout the manuscript. All decisions regarding the final wording and the incorporation of suggested translations remained with the authors. The authors take full responsibility for the accuracy, integrity, and scientific content of the published article, including any errors or omissions in the final text.

References

1. Baye MR, Morgan J, Scholten P (2004) Price dispersion in the small and in the large: Evidence from an internet price comparison site. The Journal of Industrial Economics 52(4):463–496
2. Berndt ER, Rappaport NJ (2001) Price and quality of desktop and mobile personal computers: A quarter-century historical overview. American Economic Review 91(2):268–273
3. BerriAI (2026) LiteLLM model prices and context window catalogue. GitHub, URL https://github.com/BerriAI/litellm/blob/main/model_prices_and_context_window.json, accessed 13 October 2026
4. Bommasani R, Hudson DA, Adeli E, et al (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258
5. Boulnois F (2025) llm-leaderboard-csv: Daily CSV snapshots of the LMArena and Open LLM leaderboards. GitHub, URL https://github.com/fboulnois/llm-leaderboard-csv, accessed 13 October 2026
6. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39(3/4):324–345
7. Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems 33:1877–1901
8. Burnell R, Hao H, Conway ARA, et al (2023) Revealing the structure of language model capabilities. arXiv preprint arXiv:2306.10062
9. Cameron AC, Gelbach JB, Miller DL (2011) Robust inference with multiway clustering. Journal of Business & Economic Statistics 29(2):238–249

10. Chiang WL, Zheng L, Sheng Y, et al (2024) Chatbot Arena: An open platform for evaluating LLMs by human preference. In: Proceedings of the 41st International Conference on Machine Learning, pp 8359–8388
11. Chitra T (2026) The souq: Value creation and extraction in openrouter. Robot Ventures, URL https://robvc.com/research/the-souq, accessed 13 October 2026
12. Cliff N (1993) Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin 114(3):494–509
13. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334
14. Elo AE (1978) The rating of chessplayers, past and present. Arco Publishing
15. Fedus W, Zoph B, Shazeer N (2022) Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research 23(120):1–39
16. Fourrier C, Habib N, Lozovskaya A, et al (2024) Open LLM Leaderboard v2. Hugging Face, URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, accessed 13 October 2026
17. Griliches Z (1961) Hedonic price indexes for automobiles: An econometric analysis of quality change. In: The price statistics of the federal government. National Bureau of Economic Research, pp 173–196
18. Hendrycks D, Burns C, Basart S, et al (2021a) Measuring massive multitask language understanding. In: International Conference on Learning Representations
19. Hendrycks D, Burns C, Kadavath S, et al (2021b) Measuring mathematical problem solving with the math dataset. In: Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)
20. Herbrich R, Minka T, Graepel T (2007) TrueSkill: A Bayesian skill rating system. Advances in Neural Information Processing Systems 19:569–576
21. Hirschman AO (1964) The paternity of an index. The American Economic Review 54(5):761–762
22. Ho A, Besiroglu T, Erdil E, et al (2024) Algorithmic progress in language models. arXiv preprint arXiv:2403.05812
23. Hoffmann J, Borgeaud S, Mensch A, et al (2022) Training compute-optimal large language models. arXiv preprint arXiv:2203.15556
24. Hunter JD (2007) Matplotlib: A 2d graphics environment. Computing in Science & Engineering 9(3):90–95
25. Ilić D, Gignac GE (2024) Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? Intelligence 106:101858
26. Jolliffe IT, Cadima J (2016) Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A 374(2065):20150202
27. Kaplan J, McCandlish S, Henighan T, et al (2020) Scaling laws for neural language models. arXiv preprint arXiv:2001.08361
28. Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50
29. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47(260):583–621

30. Liang P, Bommasani R, Lee T, et al (2023) Holistic evaluation of language models. Transactions on Machine Learning Research
31. LMArena (2025a) Leaderboard changelog. arena.ai, URL https://arena.ai/company/leaderboard-changelog, accessed 17 October 2026
32. LMArena (2025b) Starting next week, Style Control will be the default view on the LMArena leaderboard. X, URL https://x.com/arena/status/1923398953468678529, post of 16 May 2025; Accessed 17 October 2026
33. LMSYS (2025) FastChat: rating_systems.py (Bradley–Terry rating computation with an anchored baseline). GitHub, URL https://github.com/lm-sys/FastChat/blob/main/fastchat/serve/monitor/rating_systems.py, accessed 17 October 2026
34. Long JS, Ervin LH (2000) Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician 54(3):217–224
35. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18(1):50–60
36. Maslej N, Fattorini L, Perrault R, et al (2025) The AI Index 2025 annual report. Stanford Institute for Human-Centered AI
37. McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp 56–61
38. Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708
39. Nordhaus WD (2007) Two centuries of productivity growth in computing. The Journal of Economic History 67(1):128–159
40. O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Quality & Quantity 41(5):673–690
41. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830
42. Rein D, Hou BL, Stickland AC, et al (2024) GPQA: A graduate-level Google-proof Q&A benchmark. In: Conference on Language Modeling (COLM)
43. Rosen S (1974) Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82(1):34–55
44. Ruan Y, Maddison CJ, Hashimoto T (2024) Observational scaling laws and the predictability of language model performance. Advances in Neural Information Processing Systems 37
45. Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems 36:55565–55581
46. Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference, pp 92–96
47. Spearman C (1904) The proof and measurement of association between two things. The American Journal of Psychology 15(1):72–101
48. Sprague Z, Ye X, Bostrom K, et al (2024) Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In: International Conference on Learning Representations

49. Srivastava A, Rastogi A, Rao A, et al (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research
50. Stigler GJ (1961) The economics of information. Journal of Political Economy 69(3):213–225
51. Strathern M (1997) 'Improving ratings': Audit in the British university system. European Review 5(3):305–321
52. Suzgun M, Scales N, Schärli N, et al (2023) Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In: Findings of the Association for Computational Linguistics: ACL 2023, pp 13003–13051
53. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in Neural Information Processing Systems 30:5998–6008
54. Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17:261–272
55. Wang Y, Ma X, Zhang G, et al (2024) MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems 37
56. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1(6):80–83
57. Wortsman M, Ilharco G, Gadre SY, et al (2022) Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Proceedings of the 39th International Conference on Machine Learning, pp 23965–23998
58. Zheng L, Chiang WL, Sheng Y, et al (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36:46595–46623
59. Zhou J, Lu T, Mishra S, et al (2023) Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911