THE MODELS BEHIND THE CLAIM: EMPIRICAL STANDARDS FOR LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting an evaluation panel is a consequential but largely unguided part of Large Language Model (LLM) research: we characterize the primary approaches researchers can use to determine how many and which models to include in an LLM evaluation, and we examine the advantages and limitations of each approach, including global standards derived from freshly published, year-long research on arXiv and local standards within the ICLR community. To see how these choices are made in practice, we trace model use across 11,428 arXiv papers that report LLM evaluations and compare this broader record with 486 eligible papers from the ICLR'26. Across this record, a consistent pattern emerges: most evaluations rely on a small set of models drawn from even fewer model families, so evaluation results may only represent a narrow portion of the LLM landscape; the median paper tests 3 models from 2 families on one task family, and 39.1% test a only single family. The ICLR comparison bolsters this point: papers associated with the venue tend to include more model variants, with most of the additional breadth coming from releases within the same families. Collectively, no single minimum model count can determine whether a panel is adequate; adequacy depends on what a paper claims, what it evaluates, how current its evidence is, and which model populations it intends to represent. We therefore translate these patterns into condition-specific standards for panel breadth, family and task coverage, recency, and model access. The resulting framework gives researchers a practical basis for planning LLM evaluation panels and gives the community a clearer basis for judging whether the evidence supports the claims.

1 INTRODUCTION

An empirical claim about a language-model method is bounded by the models used to test it (Raji et al., 2021). Evaluating three sizes of one lineage can reveal sensitivity to scale within that lineage (Biderman et al., 2023). Comparing model series with documented development differences can reveal sensitivity to training data, architecture, or alignment (Liang et al., 2023). Testing a current hosted model can establish relevance to deployed systems, while testing an open checkpoint can support controlled inspection and replication (Groeneveld et al., 2024). These panels may have the same number of rows in a results table, but they provide different evidence.

This distinction has become more consequential as the model ecosystem has expanded (Stanford Institute for Human-Centered Artificial Intelligence, 2026). Authors now choose among families, parameter scales, base and instruction-tuned variants, access conditions, modalities, and dated or continuously updated services (Laskar et al., 2024). The resulting comparison set often appears as implementation detail, even though changing that set can change the apparent strength and scope of a result (Perlitz et al., 2024). Work on benchmark choice, prompt sensitivity, and underspecification has established that evaluation conclusions depend on design decisions beyond the method being evaluated (Dehghani et al., 2021). Model selection is one of those decisions.

Existing surveys organize what Large Language Model (LLM) evaluations measure and how they measure it (Laskar et al., 2024). Empirical studies have mapped metrics, benchmarks, and reporting practice (Yang et al., 2026). These accounts provide essential taxonomies, while recent studies also synthesize reported model performance (Park et al., 2025) and audit its distance from the contemporary frontier (Gringras & Salahshoor, 2026); our question is how model, family, task,

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.

access, and temporal coverage jointly inform the design of an evaluation panel. Mention counts cannot answer this question. A model named in related work, copied from another paper's leaderboard, used only to generate data, or used only to judge outputs does not supply evidence about that model's task performance in the source paper. Nor is there one universally correct way to choose a panel. A bounded deployment study may evaluate every candidate system. A method paper may need a small set of carefully matched controls. A broad capability claim may require coverage across families, access conditions, scales, and release periods. Researchers also rely on canonical baselines, leaderboards, stability checks, feasibility limits, and precedent in the literature. Each approach answers a different question and can mislead when treated as a general rule. A model count is therefore meaningful only after the rationale for choosing the models has been made explicit.

We address this gap with a freshly-published one-year census of arXiv. The population includes every unique record first submitted from 15 September 2025 through 14 September 2026, across all categories. We extract author-executed model evaluations and preserve two levels of identity: the most specific reported model and its broader family. We also classify the measured task using a two-level taxonomy, relate exact releases to their public release dates, and synthesize reported outcomes only within recovered score columns with an identifiable task and metric direction.

The study asks four questions:

1. Which approaches can justify the composition of an LLM evaluation panel in theory, and what are their respective advantages and limitations?
2. What model-family, access, and temporal coverage do current LLM evaluations provide?
3. How do the synthesized empirical standards differ across various evaluation tasks and research contributions?
4. Which quantitative standards should be followed for planning studies?

Our contribution is both methodological and empirical. We first synthesize seven recurring approaches to panel selection and state the decision rule, evidential value, and failure mode of each. We then examine evaluation practice at scale, characterizing which models researchers test, how evaluation panels vary across research settings, and how these choices shape the strength and scope of reported claims. The analysis establishes empirical reference points for panel breadth, model recency, and evidence coverage across contribution and task types. We use the ICLR 2026 proceedings as a focused community-level comparison, examining how evaluation practices in a selective machine-learning venue relate to patterns across the broader literature. Together, the presented framework and measurements aim to make model selection a tenable and verifiable part of experimental design in studies involving LLMs.

2 Related work

Evaluation scope and validity. HELM organizes evaluation across models, scenarios, and metrics, while BIG-bench examines capabilities across tasks and model scales (Liang et al., 2023; Srivastava et al., 2023). Surveys identify limitations in evaluation coverage and comparability (Chang et al., 2024; Laskar et al., 2024). Recent studies examine whether benchmarks measure their intended constructs and whether shared capability labels imply similar measurements (Bean et al., 2025; Desai et al., 2026). These findings motivate coverage tied to the research claim; our task families organize reported evaluations rather than establish independent capabilities.

Empirical studies of evaluation practice. Audits document metric use and changes in evaluation practice (Blagec et al., 2022; Schmidtova et al., 2024; Yang et al., 2026). LLMEvalDB synthesizes reported experiments for four proprietary models, including prompting effects and comparisons with peer-reviewed subsets (Park et al., 2025). Frontier Lag examines evaluated models, elicitation settings, and capability claims against contemporary performance frontiers (Gringras & Salahshoor, 2026). These studies establish precedents for cross-paper synthesis and temporal audits. Our contribution connects model, family, task, access, and temporal coverage to condition-specific planning references. Model age describes timing; it does not measure distance from a capability frontier.

Comparator choice and evaluation resources. Comparison design links baseline choice to the scientific question (Patterson et al., 2024). Rankings can depend on the competitor pool

(Benavoli et al., 2016; Perlitz et al., 2024), and errors remain correlated across different providers and architectures (Kim et al., 2025). Family coverage therefore does not establish behavioral independence. Active selection and best-model identification prioritize evaluations within a specified candidate pool (Durmazkeser et al., 2025; Tolochinsky et al., 2026); efficient querying and sequential methods address performance estimation over evaluation items (Li et al., 2025; Wu et al., 2026; Hsu & Shekhar, 2026). These methods complement the broader question of which models support a scientific claim. Their guarantees do not determine a sufficient number of model families.

Reporting and evidence synthesis. Model cards and reproducibility guidance support explicit model identities and evaluation protocols (Mitchell et al., 2019; Pineau et al., 2021; Biderman et al., 2024). NIST AI 800-2's initial public draft connects evaluation objectives, implementation, and reporting (Center for AI Standards and Innovation, 2026); NIST AI 800-3 distinguishes tested-item performance from generalization to an item population (Keller et al., 2026). Our panel references complement these principles. Evidence mapping and transparent synthesis provide a basis for organizing heterogeneous studies (James et al., 2016; Petersen et al., 2015; Campbell et al., 2020). We retain source scores on their original scales and calculate ranks within compatible comparison blocks.

3 Selecting an Evaluation Panel

The composition of an evaluation panel shapes what a study can reasonably claim. Evidence from one model release cannot automatically support conclusions about an entire family or the wider model ecosystem (Raji et al., 2021; D'Amour et al., 2022). Researchers must therefore define the population of systems their evaluation is intended to represent (Mitchell et al., 2019; Liang et al., 2023). Although no universal formula determines the required number of models, the literature provides seven complementary approaches to this decision:

Bounded census. A bounded census evaluates every model in an explicitly defined candidate population, such as all systems available to an organization on a stated date. It provides the clearest coverage when the claim concerns that exact population. Its value depends on a defensible boundary, however, and it becomes costly and quickly outdated when the target is defined broadly as "current LLMs." Large evaluation programs therefore specify the universe they cover and acknowledge exclusions rather than implying universal coverage (Liang et al., 2023; Srivastava et al., 2023; Raji et al., 2021).

Claim-driven contrasts. Claim-driven selection includes the models needed to distinguish the focal explanation from plausible alternatives (Patterson et al., 2024). Comparing methods on one model tests the intervention, whereas repeating the comparison across models tests its scope (Biderman et al., 2023; Liang et al., 2023). The panel should therefore contain the differences in scale, developer, access, or architecture required by the claim. This approach is efficient and interpretable, but it supports only the contrasts built into the design. Weak or poorly tuned baselines can make an apparent advance disappear under re-evaluation (Dacrema et al., 2019; Patterson et al., 2024).

Stratified coverage. Stratified selection identifies relevant dimensions—such as family, access, scale, specialization, modality, and release period—and covers the combinations needed by the claim. It makes heterogeneity visible rather than treating models as interchangeable (Mitchell et al., 2019; Liang et al., 2023). Its validity depends on the taxonomy: one release may not represent an entire family, and correlated strata can make a panel appear broader than it is (Kim et al., 2025).

Precedent or leaderboard selection. This approach uses canonical baselines, widely evaluated models, or current leaderboard leaders. It supports comparison with prior work and makes replication easier, but popularity and rank reflect earlier selection decisions (Pineau et al., 2021; Singh et al., 2025). Rankings can change with the benchmark, prompt, scoring rule, and competitor pool (Dehghani et al., 2021; Mizrahi et al., 2024; Alzahrani et al., 2024; Perlitz et al., 2024). Precedent alone may therefore reinforce concentration or lag behind the release frontier (Singh et al., 2025; Gringras & Salahshoor, 2026).

Adaptive stability. Adaptive selection adds or removes models until a prespecified conclusion becomes stable. Competitor composition can change benchmark rankings (Perlitz et al., 2024), but existing work on reducing evaluation examples concerns a different unit of analysis (Polo et al., 2024; Li et al., 2025). Applying stability checks to model selection is therefore a design strategy rather than an established sufficiency guarantee, and sequential addition is a form of optional stopping: a run of similar early models can end the search prematurely. The candidate pool, addition order, tolerance, and stopping rule should be specified in advance, with the final conclusion checked on reserved contrasts.

Cost-constrained design. Cost-constrained design is an operational filter applied after the evidentiary requirements of the study have been identified. When cost or access prevents the ideal panel from being evaluated, a defensible design preserves the contrasts essential to the claim and first reduces redundant prompts, examples, or closely related variants. Efficient benchmarking methods show that evaluation effort can sometimes be reduced without changing the conclusion, although seemingly minor design changes can also alter the leader (Perlitz et al., 2024; Polo et al., 2024; Wu et al., 2026). Resource limits define what can be executed; they do not by themselves justify broader generalization (Raji et al., 2021).

Empirical standards. Observed practice provides a descriptive reference for the breadth of an evaluation panel rather than an independent test of validity. Global standards describe patterns across a broad research population, while local standards describe the comparison sets used within a particular community (Schmidtova et al., 2024; Park et al., 2025). These references are transparent and attainable, but inherit historical concentration, unequal access, publication selection, and delay (Liang et al., 2023; Singh et al., 2025; Gringras & Salahshoor, 2026). They should therefore guide planning and prompt justification after the claim's evidentiary requirements have been established.

These approaches serve different roles in panel design. Bounded census, claim-driven contrasts, and stratified coverage specify the evidence needed to support the intended claim. Precedent, adaptive stability, and empirical standards guide the choice of models or help determine when expanding the panel is unnecessary. Cost constrains design choices. Appendix A further explains what each approach contributes.

4 METHODS

To connect the methodological approaches described above with current practices, we conduct an empirical study of LLM research to examine how evaluation panels are constructed in current research and use those observations to establish condition-specific references for future studies. We analyzed one-year of arXiv research, identifying the models authors actually tested, the tasks on which they tested them, and the timing of those choices. We then use these data to establish condition-specific references and examine how broader research practices compare with those of the ICLR community.

4.1 STUDY POPULATION AND ELIGIBILITY

We constructed a paper-level map of LLM evaluation practice covering arXiv records first submitted from September 15, 2025 through September 14, 2026. Starting from the complete arXiv feed, rather than an LLM-keyword search, allowed potentially relevant papers to enter the screening process regardless of how their titles or abstracts described the work. The complete census contained 337,286 original records from which screening identified 91,791 papers with evidence of possible LLM evaluation. The OAI-PMH feed (arXiv, 2026b) defines the corpus, and the arXiv API (arXiv, 2026a) resolves submission dates and document versions. We use the first submission to define the study cohort and the latest available version to recover evaluation evidence.

We identify author-executed evaluations using model-name patterns and contextual rules for tasks, execution, and outcome attribution with a deterministic rule-based extractor. These rules exclude detected mentions, auxiliary uses, and copied results. Eligible systems are decoder-based, encoder–decoder, recurrent, or multi-modal generative language models; encoder-only masked models, embedding models, and rerankers are outside scope. Each instance forms a *paper–model–task relation*. The corpus contains 65,863 relations from 11,428 papers, each linked to its supporting passage and

normalized labels. Appendix B reconciles the counts; Appendices C–K document eligibility and sensitivity analyses.

Each record is represented by exact model name, release-specific identifier, and operational family. These levels respectively support reproducibility, temporal comparison, and breadth across related series (operational families do not establish independent training ancestry; Appendix E). Benchmark mappings and experimental context assign 12 broad task families for stable comparisons and 45 operational classes for specificity. Contribution labels are non-exclusive because papers often serve several roles. The release registry records date, product track, and access class. Model age runs from release to first submission; generic labels contribute to breadth but not release-specific analyses.

4.2 Empirical standards

Panel breadth. We measure breadth by the number of distinct models, model families, and task families. The *upper-quartile standard*, R_{75}, is the observed 75th percentile for each dimension within a contribution group. A panel meeting this standard is at least as broad as those in approximately three quarters of comparable papers. The upper quartile represents strong but routinely observed practice without being driven by exceptionally large panels. We also estimate model and family standards by task and report joint attainment in Appendix L.2. We compare whole-paper and within-task family coverage among papers evaluating multiple tasks, weighting each paper equally (Appendix F).

Model recency. For each paper and task t, we measure the age of the newest model evaluated on that task from its release date to the paper's first submission. The *recency standard*, $A_{25}(t)$, is the observed 25th percentile of these ages and therefore represents the most current quarter of comparable papers. When a reliable task-specific estimate is unavailable, we use the corpus-wide threshold of 170 days, calculated from the newest model evaluated in each paper. Using first submission fixes the temporal reference, although later revisions can change the recovered panel. Breadth uses all eligible papers in the group; recency uses those with a nonnegative model age on the relevant task. Their 95% confidence intervals are obtained from 2,000 paper-level bootstrap resamples.

We also recalculate breadth references in four consecutive three-month submission cohorts, overall and within contribution groups (Appendix F.3). This tests sensitivity to the submission period using the recovered versions.

A supplementary analysis examines the sensitivity of cross-paper performance summaries to comparator choice and resampling (Appendix H).

4.3 ICLR community reference

We use ICLR 2026 as a community-specific reference. Conference appearance is descriptive, not a quality label, because review outcomes vary. Exact normalized-title matching links 486 evaluation-eligible ICLR papers to the census(ICLR, 2026). Because venue-linked papers differ in timing and research composition, ridge-logistic overlap weights balance month, arXiv category, tasks, and contribution while emphasizing comparable papers (Li et al., 2018); 2,000 group-stratified bootstraps refit the weights. Since task count is built from the task indicators, a second specification omits them when estimating task breadth. Neither specification estimates a causal effect on acceptance (Appendix I).

5 Results

5.1 Most evaluation panels remain narrow and concentrated

Most LLM evaluations rely on a small and closely related set of models. The median paper tests 3 distinct models from 2 families on a single task family; 39.1% evaluate one family, and 63.6% evaluate no more than two. For individual tasks, 46.3% of paper–task units draw all of their evidence from one family. Figure 1 summarizes this concentration across panel-size thresholds, whole-paper and within-task comparisons, and task-level family breadth. Additional rows often add releases or sizes from a represented family. Of 2,873 papers testing at least five models, 12.6% cover no more than two families. Alibaba Qwen, OpenAI GPT, Meta Llama, Google Gemini, and Anthropic Claude

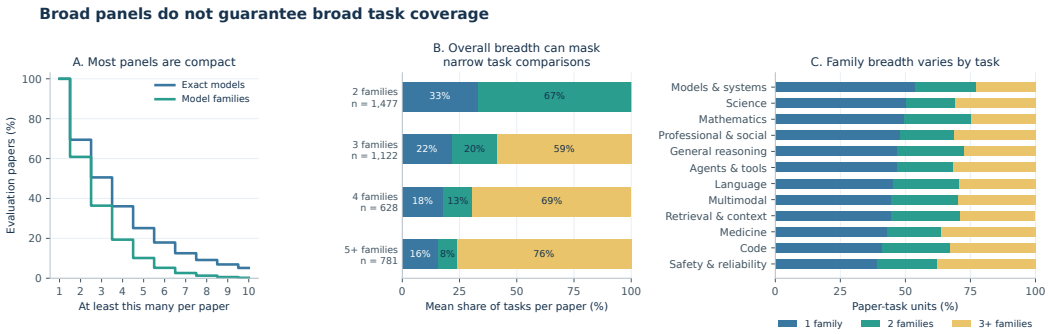


Figure 1: Evaluation breadth and task coverage. A: successive breadth thresholds. B: mean within-paper task shares by whole-paper family count, restricted to papers with multiple tasks and families; n counts papers. C: family counts across all paper–task units by task. Colors in B and C denote one, two, or at least three families.

account for 66.9% of paper–family records. GPT-4o is the most frequently tested model (1,497 papers), followed by Qwen3 (857) and GPT-5 (746). (Appendices F and E).

Whole-paper breadth did not consistently carry through to individual tasks. Among 2,531 papers evaluating multiple tasks and at least three families overall, 1,042 (41.2%) evaluated at least one task using a single family (Appendix F).

5.2 Evaluation standards depend on the claim, task, and time

The appropriate panel standard depends on the contribution. Across the full corpus, the upper-quartile standards are 5 distinct models, 3 model families, and 2 task families. The standard for *benchmark and evaluation papers* is six models from four families. For *domain applications*, it is five models from four families across three tasks; for *systems papers*, it is four models from three families across three tasks. Figure 2 and Appendix L report the remaining contribution families.

Whole-paper breadth often masks narrower evidence for individual tasks. For 10 out of the 12 task families, the within-task upper-quartile standard is four distinct models and three families. The standard is three models and two families for *models and systems*, and four models and two families for *mathematics*. Across task families, only 21.1–32.8% of paper–task units meet both their model-count and family-count standards simultaneously.

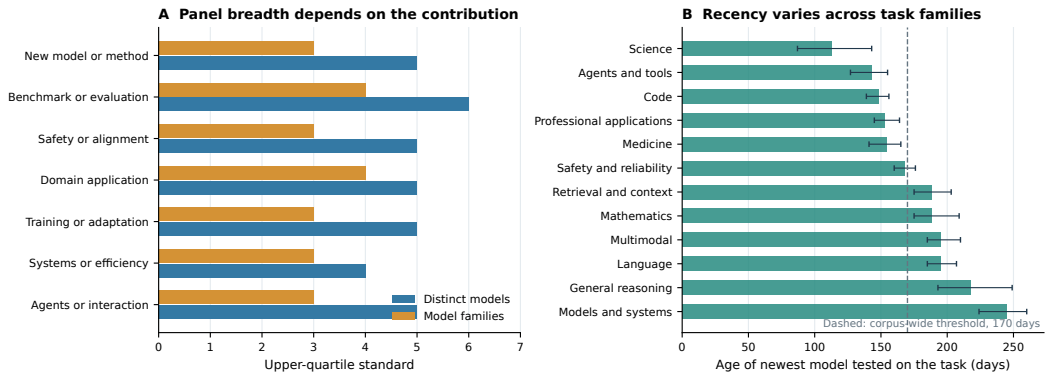


Figure 2: Condition-specific evaluation standards. Panel A reports contribution-specific upper-quartile panel breadth. Panel B shows the newest-model ages marking the most current quarter of each task family, with 95% bootstrap intervals.

Broad and current panels are uncommon. Among the most current quarter of papers, the newest model evaluated on a task is between 113 and 245 days old, depending on the task family. Across

papers overall, the corresponding threshold is 170 days, whereas the median evaluated model is 400 days old. Using six months (180 days) as a rounded version of the overall threshold, only 17.0% of release-matched papers combine a model released within that period with coverage of at least three families.

Breadth, access, and recency rarely coincide in the same evaluation. Only 14.5% of papers simultaneously meet the overall standards of five models, three families, and two tasks. Within individual contribution types, joint attainment ranges from 9.6% to 17.9%. Requiring both open-weight and hosted models and at least one model released within 170 days reduces this range to 2.0–8.7%. Each added condition applies only when the claim requires it (Appendix L.2).

Consistency across submission cohorts. The overall family reference remains three in all four cohorts, while the model reference is four or five. Across the 28 contribution–cohort combinations, family references match the full-year values in 26, compared with 19 for models and 20 for tasks; all differences are one unit. Benchmark papers retain references of six models and four families throughout, whereas safety papers range from four to six models and three to four families. Family breadth is therefore more consistent across these cohorts than exact panel size, but contribution and task references still need periodic updating (Appendix F.3).

Comparator sensitivity. In the retained comparison subset, 19.3% of 2,005 non-tied pairwise orderings reversed between within-panel percentiles and a comparator-adjusted Bradley–Terry model, and neither method reliably predicted held-out orderings, so cross-paper comparisons need shared, task-relevant baselines (Appendix H).

5.3 ICLR papers add variants more often than families

ICLR papers use longer panels mainly by adding variants within represented families. After balancing time, category, contribution, and task composition, ICLR papers contain 0.53 more distinct models per paper (95% CI [0.21, 0.87]) but only 0.13 more families ([−0.02, 0.29]); within-family additions account for 76% of the model-count difference. Allowing task composition to vary produces the same pattern: differences of 0.74 models, 0.21 families, and 0.21 tasks, with 72% of the model-count difference occurring within families. ICLR papers are also 6.2 percentage points more likely to span open-weight and hosted models. The ICLR comparison therefore shows greater panel depth and access coverage, while the family contrast is smaller and sensitive to the adjustment specification (Appendix I).

6 Implications and Standards for LLM Evaluation

6.1 Standards should follow the claim

The central implication is that an evaluation panel has several distinct dimensions. Exact identifiers measure the number of rows; families measure coverage across model series; tasks determine where the evidence applies; release age determines temporal relevance; and access determines whether the panel represents inspectable checkpoints, deployed services, or both. A single minimum model count cannot substitute for these dimensions.

The selection approach in Section 3 should be chosen before consulting a numeric reference. A bounded census requires no model-count threshold because its adequacy follows from population coverage. A claim-driven study is judged by whether every necessary control is present. A stratified or current-capability study, however, benefits from the family, task, access, and recency references derived here. Cost-constrained and adaptive designs should report which reference they could not meet and demonstrate that the focal conclusion remains stable under feasible alternatives. Under tight budgets, we suggest securing the required contrasts first, then focal-task family coverage, then a recent model if current capability is claimed, and only then more releases within represented families.

Table 1 uses separate upper-quartile references because each dimension answers a different design question. Additional releases test variation within a series, additional families broaden series coverage, and additional tasks test transfer; keeping these references separate makes omissions visible and prevents a large count on one dimension from substituting for another. Across four narrower

Table 1: Claim-dependent planning and reporting standards. Count thresholds are separate marginal references, not a joint upper-quartile profile. Apply access and recency rows when the claim requires them. $A_{25}(t)$ uses models evaluated on task t. The three-family recommendation exceeds the observed two-family reference for systems and mathematics.

Claim or use	Planning or reporting reference	Application
Contribution profile	Identifiers/families/tasks: new model, training, safety, or agent work, 5/3/2; benchmark work, 6/4/2; domain applications, 5/4/3; systems work, 4/3/3.	Compare each relevant dimension with its marginal R_{75}. Joint attainment is reported in Appendix L.2.
Cross-family coverage	Use 3 operational families as the planning reference on each focal task.	For generalization, justify relevant differences in training or architecture; family names alone do not establish independence.
Current capability	Include a model evaluated on the focal task no older than $A_{25}(t)$; use 170 days as the overall fallback.	Record the panel-selection date; empirical ages use first submission. Historical studies use their stated time boundary.
Cross-access relevance	At least one open-weight and one hosted model when the claim spans both conditions.	The two access classes support different claims about inspection, reproduction, and deployed capability.
Minimum reporting	Release or snapshot, family, developer and known base model, access, scale when known, evaluation date, and model-by-task coverage.	These fields make model relationships, timing, and task coverage auditable.

extraction screens, the overall family upper-quartile standard remains three while the exact model upper-quartile standard is four or five (Appendix J). The cohort analysis supports retaining family coverage as a planning reference while updating exact counts against recent papers with similar contributions. Authors apply the relevant dimensions and justify departures rather than treating a one-model difference as a validity boundary. Joint attainment rates show how demanding each combination is.

6.2 Panel standards distinguish replication from generalization

For an overall contribution, the appropriate starting point is the matching R_{75} value in Figure 2. Benchmark and domain papers have the largest family reference (four), while systems work has a smaller exact-model reference but broader task coverage. These differences are substantively coherent: a benchmark claims comparative coverage, a domain paper must test transfer beyond a familiar setting, and a systems paper often makes claims across efficiency and quality outcomes.

The focal-task reference is more informative than the whole-paper total. Ten of twelve task families have marginal references of four identifiers and three families, and 21.1–32.8% of paper–task panels meet both task-specific count thresholds. For systems and mathematics, three is a planning recommendation above the observed reference of two. A focused two-family study can justify its scope through the contrast it tests; one-family designs address release or scale sensitivity.

Model selection should combine continuity with documented contrasts in training, architecture, access, or other properties relevant to the claim. Record shared developers and base models where known; GPT and o-series provide two series but one developer.

Appendix D maps common comparison sets; model documentation establishes which differences they test. A leaderboard leader is not automatically a valid control, and three releases from one family do not establish generalization across families.

6.3 Temporal and access standards define the population of systems

A claim about present-day capability needs a dated model evaluated on the focal task. $A_{25}(t)$ provides that task-specific reference, ranging from 113 to 245 days, with a 170-day paper-wide fallback. Authors should document panel-selection, execution, and submission dates. A prolonged

protocol can justify its fixed evaluation window by bounding the claim to that date, rather than adding a nominally recent model with incomparable measurements.

Older models still have clear evidential roles. A stable checkpoint supports replication, comparison with prior work, and controlled inspection. The temporal standard therefore does not require replacing established baselines; it requires pairing them with a recent model when the claim concerns current capability. A historical replication can omit the frontier model if its conclusion is explicitly historical.

Access requires the same alignment between evidence and claim. Open-weight models support inspection and controlled deployment, while hosted models represent systems that researchers and users encounter through services. A claim spanning both populations should test at least one of each (Table 1). If the method requires weight access or a hosted feature, the paper should state that restriction rather than implying broader applicability.

6.4 The ICLR comparison informs review of evaluation coverage

The ICLR comparison in Section 5 describes venue-associated evaluation designs. Additional identifiers predominantly represent releases or sizes within existing families under both adjustment specifications. Task breadth is modestly greater when it is not directly balanced. These observations favor explicit coverage arguments over table length; they do not reveal which choices reviewers rewarded or whether models were added during review. Reviewers should assess which family, task, temporal, and access contrasts support the claim, and why the design departs from the relevant reference. A large panel cannot compensate for weak provenance, incomparable settings, or a focal task evaluated on one family. Authors should state the target population and claim, name the selection approach, assign each model an evidential role, compare the panel with the relevant global or local reference, and test feasible additions or removals.

7 Conclusion

In studies involving LLMs, model selection determines which conclusions the study can support. The presented analysis provides condition-specific references for model, family, task, access, and temporal coverage. We bring together approaches to model selection and evidence from 11,428 arXiv papers, including a comparison with the ICLR community, to develop standards for different research contributions and tasks. The findings show why model counts alone are insufficient: additional models often represent variants within existing families, and a broad panel across a paper can conceal narrow coverage on individual tasks. Model relationships, release dates, and access conditions also determine the scope of the comparison.

The proposed standards give researchers a concrete basis for choosing models and explaining how their evaluation supports its claims. The evaluation panel should center on the evidence needed for each claim, with numerical thresholds informing that judgment. As models and evaluation practices change, the standards must be updated. Reporting why each model was selected, where it was tested, and which conclusions it supports should become a routine part of LLM evaluation.

8 Limitations

The standards characterize observed practice within a defined one-year arXiv corpus; they are planning anchors rather than validated minimum requirements for every evaluation. The evidence map depends on explicit, recoverable reporting, so evaluations confined to unavailable supplements, rasterized tables, or ambiguous model labels may be undercounted. Independent extraction accuracy has not been measured; confidence intervals condition on the recovered labels. Operational families represent related product series and do not necessarily imply independent developers, training data, or model ancestry. Release-based analyses are limited to models that can be matched confidently to a dated release. Persistence across years and venues is untested (Yang et al., 2026). Finally, the ICLR comparison covers one venue-year, uses exact normalized-title matches, and is observational and cannot show that a panel affected acceptance. These boundaries limit universal or causal interpretations while preserving the standards' use as transparent, updateable descriptions of current practice.

AI USE STATEMENT

Generative AI tools assisted with implementing analysis code, creating figures, identifying literature, and editing the manuscript. The authors take responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The supplementary material includes a register of the papers studied, extracted model–task records, model and task taxonomies, and model release information. It also provides the data and code supporting the reported analyses.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744.

arXiv. arXiv API user's manual. https://info.arxiv.org/help/api/index.html, 2026a. Accessed 2026-09-15.

arXiv. Open Archives Initiative Protocol. https://info.arxiv.org/help/oa/index.html, 2026b. Accessed 2026-09-15.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0590. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1967e0fc3aa6cbbace562f5cb8e3954e-Abstract-Datasets_and_Benchmarks_Track.html.

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should we really use post-hoc tests based on mean-ranks? *Journal of Machine Learning Research*, 17(5):1–10, 2016. URL https://www.jmlr.org/papers/v17/benavoli16a.html.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The NeurIPS 2021 consistency experiment. https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/, 2021. Accessed 2026-09-16.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. URL https://arxiv.org/abs/2405.14782.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pp. 52–63. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.nlppower-1.6.

Mhairi Campbell, Joanne E. McKenzie, Amanda Sowden, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*, 368:l6890, 2020. doi: 10.1136/bmj.l6890.

Center for AI Standards and Innovation. Practices for automated benchmark evaluations of language models. Technical Report NIST AI 800-2, National Institute of Standards and Technology, 2026. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf. Initial public draft, January 2026.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 2024. doi: 10.1145/3641289.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint arXiv:2109.09774*, 2021. doi: 10.48550/arXiv.2109.09774.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109. ACM, 2019. doi: 10.1145/3298689.3347058.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

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. URL https://arxiv.org/abs/2107.07002.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Meera Desai, Sang T. Truong, Hanna Wallach, Alex Chouldechova, A. Feder Cooper, Jean Garcia-Gathright, Daniel E. Ho, Abigail Z. Jacobs, Sanmi Koyejo, Nicholas Pangakis, and Angelina Wang. What AI benchmarks actually measure: Adapting convergent and discriminant validity to interrogate fifty-six AI benchmarks. *arXiv preprint arXiv:2609.08812*, 2026. doi: 10.48550/arXiv.2609.08812. URL https://arxiv.org/abs/2609.08812.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.04475.

Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. Active model selection for large language models. *arXiv preprint arXiv:2510.09418*, 2025. doi: 10.48550/arXiv.2510.09418. URL https://arxiv.org/abs/2510.09418.

David Gringras and Misha Salahshoor. Frontier lag: A bibliometric audit of capability misrepresentation in academic AI evaluation. *arXiv preprint arXiv:2605.04135*, 2026. doi: 10.48550/arXiv.2605.04135. URL https://arxiv.org/abs/2605.04135v2. Version 2.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9734–9745, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models. *arXiv preprint arXiv:2607.17409*, 2026. doi: 10.48550/arXiv.2607.17409. URL https://arxiv.org/abs/2607.17409.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12804–12825. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.760.

ICLR. ICLR 2026 proceedings. https://proceedings.iclr.cc/paper_files/paper/2026, 2026. Accessed 2026-09-16.

Katy L. James, Nicola P. Randall, and Nick R. Haddaway. A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, 5(7), 2016. doi: 10.1186/s13750-016-0059-6.

Drew Keller, Kweku Kwegyir-Aggrey, Ryan Steed, Anita K. Rao, Julia L. Sharp, and A. Stevie Bergman. Expanding the AI evaluation toolbox with statistical models. Technical Report NIST AI 800-3, National Institute of Standards and Technology, 2026. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4th edition, 2019.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Xiangji Huang. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13785–13816. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.764.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi: 10.1080/01621459.2016.1260466.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25bp.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhui Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. ACM, 2019. doi: 10.1145/3287560.3287596.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

Jungsoo Park, Junmo Kang, Gabriel Stanovsky, and Alan Ritter. Can LLMs help uncover insights about LLMs? a large-scale, evolving literature analysis of frontier LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pp. 20412–20433. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.998. URL https://aclanthology.org/2025.acl-long.998/.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024. URL https://www.jmlr.org/papers/v25/23-0183.html.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2519–2536. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.139.

Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64: 1–18, 2015. doi: 10.1016/j.infsof.2015.03.007.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://www.jmlr.org/papers/v22/20-303.html.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 557–583. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.inlg-main.44.

Patrícia Schmidtová, Saad Mahamood, and Ondřej Dušek. Never truly out of fashion: A retrospective look at evaluation in NLG. In *Proceedings of the 1st Symposium on Natural Language Generation Evaluations*, pp. 63–72. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.retroeval-main.8/.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/70a93f260a51123b3c0e33ecd1b4de97-Abstract-Datasets_and_Benchmarks_Track.html.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2206.04615.

Stanford Institute for Human-Centered Artificial Intelligence. The 2026 AI index report. Technical report, Stanford University, 2026. URL https://hai.stanford.edu/ai-index/2026-ai-index-report.

Elad Tolochinsky, Yaniv Tenzer, and Yaniv Romano. Valid best-model identification for LLM evaluation via low-rank factorization. *arXiv preprint arXiv:2605.10405*, 2026. doi: 10.48550/arXiv.2605.10405. URL https://arxiv.org/abs/2605.10405.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient evaluation of LLM performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026. doi: 10.48550/arXiv.2601.20251. URL https://arxiv.org/abs/2601.20251.

Jing Yang, Nils Feldhus, Salar Mohtaj, Leonhard Hennig, Qianli Wang, Eleni Metheniti, Sherzod Hakimov, Charlott Jakob, Veronika Solopova, Konrad Rieck, David Schlangen, Sebastian Möller, and Vera Schmitt. What are we measuring in NLG? a meta-analysis of evaluation trends 2020–2025. *arXiv preprint arXiv:2601.07648*, 2026. URL https://arxiv.org/abs/2601.07648.

A RESEARCH SUPPORTING THE MODEL-SELECTION APPROACHES

This appendix explains how we identified eligible papers, counted the models they evaluated, and calculated the proposed standards for model selection. Appendices B–E explain the study population, eligibility rules, and model and task categories. Appendices F–I provide the detailed results and comparisons. Appendices J and K explain how the extraction works and how different evidence requirements affect the results. Appendix L gives the calculations behind the proposed standards. Our standards describe common and more extensive evaluation practices; researchers can use them when deciding how many and which models to test. They do not establish a minimum panel size that makes every study valid.

The approaches in Section 3 draw on selected research on evaluation methods. This is a focused review, not a systematic review of all relevant publications. Table 2 separates what earlier studies established from how those findings inform model selection. The approaches address different decisions. The research claim determines which comparisons are necessary. Model categories help identify relevant differences to test. Established baselines connect a study with previous work. Adding models in stages can show whether a conclusion depends on the initial selection. Budget limits determine what is feasible. The observed standards then show how the resulting panel compares with those used in other papers. Research on algorithm comparison distinguishes the systems being compared (Demšar, 2006; Benavoli et al., 2016). For LLMs, model names, product series, developers, and pretrained base models are different units.

Table 2: Research supporting the seven approaches to selecting models. Earlier studies inform the design choices below but do not establish a model count suitable for every evaluation.

Approach	Findings from earlier research	Implication for model selection
Bounded census	HELM and BIG-bench define which models and tasks an evaluation covers (Liang et al., 2023; Srivastava et al., 2023). Benchmark scope limits what results can show (Raji et al., 2021).	Define the eligible models and a cutoff date. Testing all of them supports conclusions about that set, rather than inaccessible or future models.
Claim-driven contrasts	Controlled comparisons help distinguish a method's effect from differences in tuning or implementation (Demšar, 2006; Patterson et al., 2024; Dacrema et al., 2019; Hooker et al., 2019; Huang et al., 2024).	Include the comparison models needed to test the claim. The required comparisons determine the panel size.
Stratified coverage	Model documentation and evaluation surveys identify properties relevant to comparison (Mitchell et al., 2019; Chang et al., 2024; Laskar et al., 2024).	Select models across the families, access types, sizes, modalities, and releases relevant to the claim. Testing one model does not establish results for its entire category.
Precedent or leaderboard	Rankings change with benchmarks, prompts, scoring rules, and the models compared (Dehghani et al., 2021; Mizrahi et al., 2024; Alzahrani et al., 2024; Perlitz et al., 2024).	Reuse established baselines to connect results with earlier studies. Popularity or rank alone does not justify the complete panel.
Adaptive stability	Evaluation studies examine sensitivity to design choices and reduce the number of test examples (Perlitz et al., 2024; Polo et al., 2024).	Check whether adding models changes the conclusion. Specify the candidate models, their order, an acceptable amount of change, and a stopping rule; check the result separately.
Cost-constrained design	Broad evaluations can be costly. Selected test examples sometimes preserve estimates at lower cost (Liang et al., 2023; Perlitz et al., 2024; Polo et al., 2024).	Protect the comparisons needed for the main claim. When justified, fewer test examples may leave resources for more models. Explain how budget limits affect the conclusions.
Empirical standards	Reviews of evaluation practice document common choices, reporting gaps, and changes over time (Blagec et al., 2022; Schmidtova et al., 2024; Schmidtová et al., 2026; Yang et al., 2026; Pineau et al., 2021).	Compare the proposed design with observed practice across the field and within the relevant community. Common practice does not establish the best design.

B STUDY POPULATION AND PAPER SELECTION

We began with all 337,286 arXiv papers first submitted during the study period. This initial collection includes papers unrelated to LLM evaluation. Recorded submission times range from September 15, 2025 at 00:00:37 UTC to September 14, 2026 at 18:33:27 UTC. We use the first submission date to determine inclusion, although the text available for analysis may come from a later revision. Different arXiv identifiers are counted as separate papers, even when they describe related work.

We screened every paper's title, abstract, and full text. Full text was available for 337,220 papers; the remaining 66 had no recoverable source or PDF text. In 25,153 papers, terms indicating possible LLM evaluation appeared in the body text but not in the title or abstract.

After screening, a rule-based program searches the available paper text for model names, task or benchmark descriptions, and statements or results indicating that a model was evaluated. It reads tables and prose using model-name dictionaries, text patterns, and the surrounding context. We refer to this process as *extraction*. Appendix K describes the extraction rules and their limitations in detail.

Within the study population, 91,791 papers have extraction entries. These are candidate papers processed to identify model evaluations, not the final number of eligible papers. Extraction produced 147,876 evidence records: entries recovered from tables, numerical lines, or prose that could support an evaluation. A single paper can produce many such records, including repeated descriptions of the same model and task. The primary inclusion rules retained 81,615 evidence records. We then assigned their task labels and combined repeated records referring to the same model on the same task in the same paper. This produced 65,863 distinct paper–model–task records from 11,428 papers. Table 3 explains the number counted at each stage.

Table 3: Paper selection and extraction counts.

Count	Unit counted	What the number represents
337,286	Papers	All papers in the initial collection, including papers unrelated to LLM evaluation.
337,220	Papers	Papers with source or PDF text available for full-text screening. Text was unavailable for the remaining 66 papers.
91,791	Papers	Candidate papers within the study period that have extraction entries. Processing a paper does not necessarily identify an eligible evaluation.
147,876	Evidence records	Entries extracted from tables, numerical lines, or prose before applying the primary inclusion rules. One paper can produce many entries, including repeated descriptions of the same evaluation.
81,615	Evidence records	Entries retained by the primary inclusion rules, before assigning all supported task labels and combining duplicate model–task records.
65,863	Model–task records	Distinct combinations of a paper, a reported model identifier, and an assigned task. Repeated evidence for the same combination is counted once.
11,428	Papers	Papers contributing at least one included model–task record. These papers form the final evaluation sample.

The 65,863 records are not 65,863 different models or independent experimental runs. Each record identifies one reported model evaluated on one assigned task in one paper. For example, a paper evaluating two models on both coding and mathematics contributes four model–task records but only one paper. A paper repeating those results in another table does not add to records.

Other analyses use different counting units. A paper–model count includes each model once per paper, regardless of how many tasks it performs. A paper–family count includes each model family once per paper. A paper–task count groups the models evaluated on the same task within a paper. For performance analysis, a comparison block is one extracted score column with an assigned task and a known direction of improvement, such as higher accuracy or lower error.

The full extraction register contains 92,687 identifiers, of which 896 fall outside the study population. Removing those leaves the 91,791 candidate papers reported above. The final screening labels identify a slightly different set of 91,615 papers with strong or moderate indications of relevant evaluation content. Three of these lack extraction entries, while 179 papers with extraction entries fall outside that screening group. This check accounts for differences between the screening and extraction lists.

C RULES FOR INCLUDING EVALUATIONS

We aim to include a model when the paper's authors tested it on a defined task and reported an outcome for that test. Mentioning a model is not sufficient. Table 4 states the inclusion rules.

Table 4: Rules for deciding whether a model evaluation belongs in the study.

Reported use	Inclusion decision	Reason
Model in a results table	Include if the authors ran it and its task and outcome are identifiable	A table can contain scores copied from elsewhere
Evaluation using cited evaluation software	Include if the authors ran the model and report its outcome	Citing software does not imply copied results
Score copied from another paper or leaderboard	Exclude	These authors did not run the evaluation
Model used only to judge another system	Exclude from the tested models; record its supporting role	The result measures the system being judged
Judge model tested for agreement or reliability	Include for that task	The result now measures the judge
Model used only to generate training data	Exclude; record its supporting role	The generator's performance is not evaluated
Language model inside a multimodal system	Count separately only if separately tested	A result for the complete system does not separately evaluate its components
Family name with a task result but no release name	Retain the name and mark the release as unknown	The result supports inclusion but not a release date
Method or prompt tested with several model APIs	Include each model the authors tested	Each model has its own task outcome
Qualitative study with a structured success/failure assessment	Include if the task and model outcome are identifiable	An outcome need not be numerical
Mention in related work or a proposal for future work	Exclude	No novel evaluation was performed

A fine-tuned model can qualify even when the paper does not test its original base model. A multimodal system counts as the system evaluated; its underlying language model counts separately only if separately tested. If a paper uses both a general family name and a compatible, more specific model name, we retain the specific name to avoid counting the same model twice. Otherwise, we retain the general name and mark its release as unknown. Model counts therefore include some reported names that do not identify a precise checkpoint.

The name dictionary covers generative decoder, encoder–decoder, recurrent, and multimodal language models, including T5 and RWKV. Encoder-only masked language models, such as BERT and RoBERTa, are outside the study's scope. Names explicitly identifying embedding-only or reranking products are also excluded. We impose no minimum parameter count. Models whose names are absent from the dictionary or unrecognized by its rules can be missed even if they otherwise meet the inclusion criteria. Appendix J shows how the results change when the evidence requirements are broadened or tightened. These checks assess the consequences of the rules.

D TASK CATEGORIES AND MODELS EVALUATED WITHIN EACH TASK

We group tasks into twelve broad families and 45 more specific categories. The dictionary also assigns 122 named benchmarks to these categories. Table 5 lists the complete hierarchy. Broad families allow comparisons across papers that use different benchmarks; the more specific categories retain distinctions within each family. A paper can evaluate several tasks, so adding the task-specific paper counts would count some papers more than once.

The intended task label describes what the evaluation measures, rather than the paper's general topic. For example, summarization in a medical paper is a language task unless the outcome measures clinical performance. A coding agent receives both code and agent labels only when separate outcomes support both assignments. Results that combine tasks without a clear interpretation are excluded from the performance analysis.

Table 5: The twelve task families and their 45 more specific categories. The released dictionary defines the categories and assigns 122 named benchmarks to them.

Broad task family	Second-level classes
Models and systems	Language modeling; training and adaptation; efficiency and serving; compression and quantization
Language and communication	Question answering and reading; summarization and text generation; translation and multilingual language; dialogue and instruction following; information extraction and classification
Safety, reliability, and society	Factuality and hallucination; safety and harmfulness; bias and fairness; privacy and security; interpretability and behavior
Multimodal perception and generation	Image understanding; document and chart understanding; video understanding; audio and speech; multimodal generation
Mathematics and formal reasoning	Mathematical problem solving; formal logic and theorem proving
Knowledge and general reasoning	General knowledge and examinations; commonsense and logical reasoning; planning and decision making
Code and software engineering	Code generation; software engineering; structured data and queries
Retrieval, memory, and long context	Retrieval and RAG; long-context processing; memory and personalization
Agents, tools, and robotics	Tool and API use; web and computer use; robotics and embodied action; games and simulations
Medicine and health	Clinical reasoning; biomedical knowledge and evidence; health communication
Science and discovery	Scientific reasoning; biology and chemistry; physics, materials, and Earth science
Professional and social applications	Law; finance and business; education; social science and humanities; creative and media work

The extraction first looks for a recognized benchmark name. Otherwise, it uses task terms in captions, nearby table text, or evaluation passages. For prose, the search can extend to the paragraph and the paper's title and abstract. Tables also use the title and abstract when the nearby text provides no task assignment. A benchmark attached to a score identifies its task more directly than a term elsewhere in the paper. Appendix J tests how results change when task labels must be supported by nearby text.

D.1 Counting all models in a paper versus models on one task

Table 6 counts all models in papers that evaluate each task family. It includes models used only on other tasks in those papers. The median is three models for most task groups. Table 7 instead counts only models assigned to the stated task. The median becomes two models for most task groups. Thus, the total number of models in a paper can overstate the models support its findings on a single task.

D.2 Which model families are used for which tasks

Figure 3 shows both widely used families and task-specific choices. Qwen, GPT, and Llama appear in all twelve task families. Unsurprisingly, LLaVA and InternVL have their largest paper counts in multimodal evaluation, while Claude is among the three most frequently used families for code. These patterns show which comparisons are common in the literature. They do not identify the best models for every future study.

Table 6: Counts of all models and families in papers evaluating each task. Models used only on other tasks in the same paper are included. CLES is the probability that a randomly selected paper in the row has more reported model identifiers than a randomly selected paper outside that row; ties count as one half.

Task family	Papers	Whole-paper models	Whole-paper families	CLES vs. other papers
Agents, tools, and robotics	1,144	3	2	0.56
Code and software engineering	1,372	3	2	0.57
Language and communication	2,651	3	2	0.58
Mathematics and formal reasoning	1,726	3	2	0.55
Medicine and health	1,583	3	2	0.58
Multimodal perception and generation	1,779	3	2	0.52
Professional and social applications	1,707	3	2	0.55
Knowledge and general reasoning	1,154	4	2	0.62
Retrieval, memory, and long context	1,426	3	2	0.56
Science and discovery	479	3	2	0.52
Models and systems	3,168	3	2	0.52
Safety, reliability, and society	2,236	3	3	0.61

Table 7: Models and families evaluated on each task. Counts are medians, with the middle 50% of observations (the interquartile range, or IQR) in brackets. *One family* gives the percentage of paper–task combinations that include only one model family. The final column lists the three most frequently used families, with shortened names.

Task family	Papers	Models	Families	One family	Frequent families
Models and systems	3,168	2 [1–3]	1 [1–2]	53.7%	Qwen / Llama / GPT
Language and communication	2,651	2 [1–4]	2 [1–3]	45.3%	Qwen / GPT / Llama
Safety, reliability, and society	2,236	2 [1–4]	2 [1–3]	39.4%	GPT / Qwen / Llama
Multimodal perception and generation	1,779	2 [1–4]	2 [1–3]	44.7%	Qwen / GPT / LLaVA
Mathematics and formal reasoning	1,726	2 [1–4]	2 [1–2]	49.8%	Qwen / Llama / GPT
Professional and social applications	1,707	2 [1–4]	2 [1–3]	48.2%	GPT / Qwen / Llama
Medicine and health	1,583	2 [1–4]	2 [1–3]	43.1%	GPT / Qwen / Llama
Retrieval, memory, and long context	1,426	2 [1–4]	2 [1–3]	44.5%	Qwen / GPT / Llama
Code and software engineering	1,372	2 [1–4]	2 [1–3]	41.3%	GPT / Qwen / Claude
Knowledge and general reasoning	1,154	2 [1–4]	2 [1–3]	47.0%	Qwen / Llama / GPT
Agents, tools, and robotics	1,144	2 [1–4]	2 [1–3]	46.9%	Qwen / GPT / Gemini
Science and discovery	479	2 [1–4]	1 [1–3]	50.5%	GPT / Qwen / Gemini

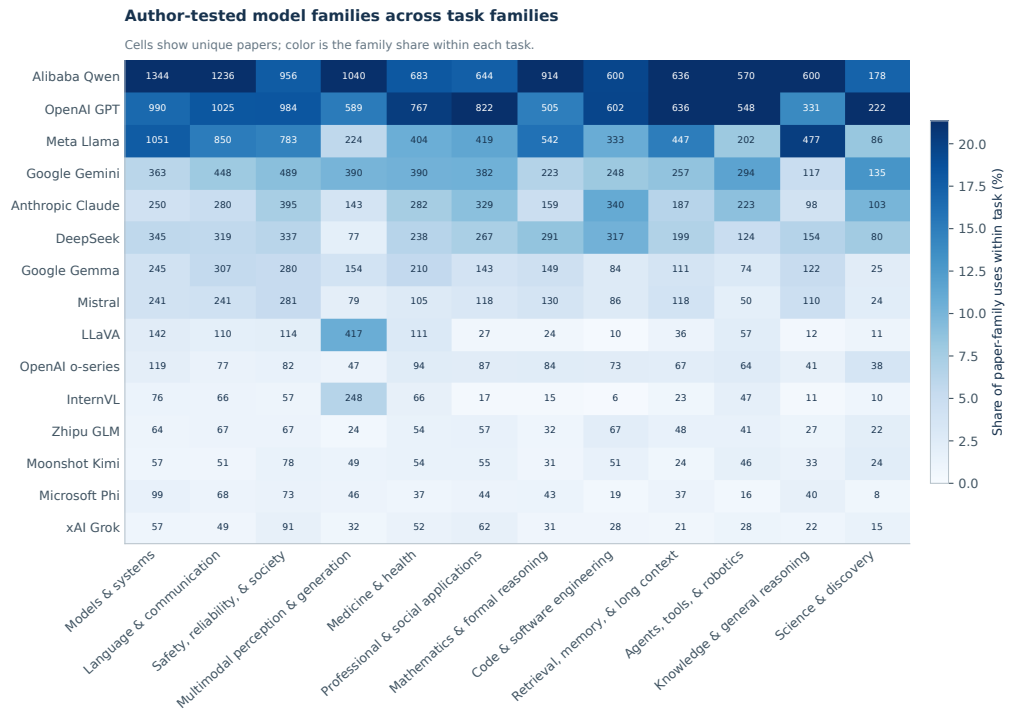


Figure 3: Model families evaluated on each task. Each number counts papers evaluating that model family on that task. Color shows its share of all paper–family records for the task, counting each model family once per paper.

E Model families, frequency of use, and model size

E.1 What we mean by a model family

We use *model family* to mean a related model series, rather than all models made by one company. Our main classification separates OpenAI GPT, o-series, and Codex, and separates Google Gemini, Gemma, PaLM, and T5. It also treats complete systems such as LLaVA and InternVL separately from their underlying language models when the reported result evaluates the complete system. The released dictionary records alternative names and the rules used to distinguish these groups.

This choice preserves differences between series from the same developer. GPT and o-series are named and released as distinct series. Gemini and Gemma also differ in access: one is a hosted model series and the other provides openly available checkpoints.[1] Keeping these series separate allows us to compare their use, access types, and release histories. We do not assume that models were trained independently. Gemma shares research and technology with Gemini. Models from different teams can also share a base model: Vicuna-13B, for example, was fine-tuned from LLaMA.[2] Brands also differ in how broadly they define a series. Family counts therefore measure coverage of the series in our dictionary, not the number of distinct training datasets, tokenizers, or alignment methods. When these differences matter to a generalization claim, they should be documented separately.

We repeated the family counts after combining GPT, o-series, and Codex into one OpenAI group, and Gemini, Gemma, PaLM, and T5 into one Google group. All other assignments stayed the same. Counts decreased in 665 of 11,428 papers (5.8%). The percentage covering at least three groups fell from 36.4% to 35.6%. The overall 75th-percentile count remained three. It also stayed unchanged in six contribution groups; for domain applications, it fell from four to three. This check tests these two

[1] OpenAI's o1 announcement: https://openai.com/index/introducing-openai-o1-preview/. Gemma Team's technical report: https://arxiv.org/abs/2403.08295.

[2] Vicuna Team's training description: https://www.lmsys.org/blog/2023-03-30-vicuna/.

combinations, rather than replacing the whole classification with developer groups. The proposed standards continue to use model series.

E.2 How often families and individual models are evaluated

Table 8 counts each family once per paper. Qwen appears in 48.7% of evaluation papers, GPT in 42.3%, and Llama in 31.1%. These percentages overlap because a paper can evaluate several families. The number of task categories records how broadly a family has been used; it does not mean that the evidence is equally extensive in every category. Table 9 reports both paper counts and model–task counts. For example, GPT-4o has 2,566 model–task records in 1,497 papers.

Table 8: The twelve most frequently evaluated model families. A family counts once per paper. Percentages can sum to more than 100% because a paper may include several families.

Model family	Papers	Share of evaluation papers	Task classes
Alibaba Qwen	5,567	48.7%	45
OpenAI GPT	4,839	42.3%	45
Meta Llama	3,556	31.1%	45
Google Gemini	2,340	20.5%	45
Anthropic Claude	1,810	15.8%	45
DeepSeek	1,810	15.8%	45
Google Gemma	1,160	10.2%	45
Mistral	998	8.7%	45
LLaVA	637	5.6%	41
OpenAI o-series	575	5.0%	42
InternVL	381	3.3%	41
Zhipu GLM	364	3.2%	45

Table 9: The fifteen most frequent reported model names. Paper counts measure how many papers evaluate each name; paper–model–task counts also distinguish tasks within a paper. A general family name is retained when a more specific release cannot be identified.

Reported identifier	Papers	Paper–model–task records	Family
GPT-4o	1,497	2,566	OpenAI GPT
Qwen3	857	1,365	Alibaba Qwen
GPT-5	746	1,275	OpenAI GPT
GPT-4o-mini	624	1,030	OpenAI GPT
Llama-3.1-8B	610	1,111	Meta Llama
Qwen3-8B	610	1,092	Alibaba Qwen
LLaVA	608	1,110	LLaVA
Gemini-2.5-Pro	559	926	Google Gemini
GPT-4	481	823	OpenAI GPT
Qwen2.5	476	753	Alibaba Qwen
Gemini-2.5-Flash	470	820	Google Gemini
DeepSeek-R1	458	724	DeepSeek
Llama-3.1-8B-Instruct	449	741	Meta Llama
Llama	436	680	Meta Llama
GPT-4.1	427	747	OpenAI GPT

We also measure how strongly use is concentrated in a few families. Let s_f be family f's share of all paper–family records. The Herfindahl–Hirschman index is $\mathrm{HHI} = \sum_f s_f^2$. Its reciprocal, $1/\mathrm{HHI}$, expresses the same concentration as a number of equally frequent families. Here, HHI is 0.113, equivalent to 8.8 equally frequent families. A second measure, the Shannon effective number, is $\exp(-\sum_f s_f \log s_f)$ and equals 13.8. These measures summarize how unevenly families are used, not their quality.

The most frequent family accounts for 20.6% of paper–family records, the top five for 66.9%, the top ten for 86.1%, and the top twenty for 95.4%. These percentages use all family records as the denominator and sum to 100% across all families. They differ from the overlapping percentages of papers in Table 8.

E.3 Model sizes that can be identified from their names

A parameter count appears in the reported name for 13,749 of 39,877 paper–model pairs (34.5%). For the remaining 65.5%, the name does not provide a size. For mixture-of-experts models, the count is the advertised total number of parameters when this appears in the name, rather than the number active for each input. Table 10 groups the known sizes. Models with at least 3 billion but fewer than 10 billion parameters form the largest named size group, accounting for 19.8% of all pairs. Because most pairs have no size identified this way, the table cannot establish the size distribution of all evaluated models.

Table 10: Model sizes identified from reported names.

Parameter band	Reported identifiers	Paper–model records	Share
$<$ 3B	193	1,699	4.3%
3B–$<$ 10B	449	7,904	19.8%
10B–$<$ 30B	130	1,315	3.3%
30B–$<$ 100B	187	2,143	5.4%
100B and above	56	688	1.7%
Not reported in identifier	997	26,128	65.5%

This is a measure of what model names reveal, not an audit of all parameter information elsewhere in the papers. When each model is counted separately for its assigned tasks, 23,082 of 65,863 records have a size label (35.0%). The proportion is similar under this alternative counting unit.

F How many models and families papers evaluate

F.1 Model counts and family counts

Figure 4 compares the number of reported model names with the number of families in each paper. Most panels are small. In Panel B, points below the diagonal represent papers that evaluate multiple models from at least one family. Adding model releases or sizes can therefore increase the model count without adding another family.

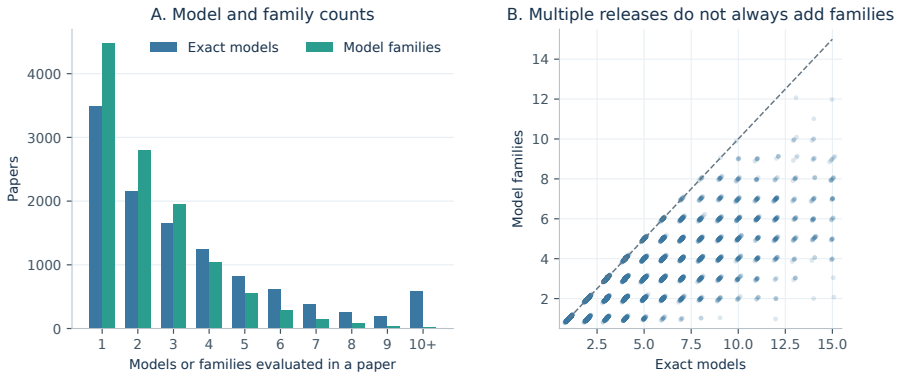


Figure 4: Model and family counts per paper. The diagonal in Panel B marks equal counts; points below it include multiple models from at least one family. Small display offsets separate overlapping points. Panel B shows the central range, while Panel A retains larger panels in its 10+ category.

The largest panel has 49 reported model identifiers from eight families. The largest family count, 12, occurs in another paper. The median paper includes 3 identifiers when general family names are retained and 2 when only names identifying a release are counted. At least one general family name with no specified release appears in 3,179 papers.

Table 11 groups papers by their research contribution, assigned from titles and abstracts. Benchmark and evaluation papers and safety or alignment papers tend to evaluate more models than training and

systems papers. A paper can belong to more than one contribution group. The effect size compares each group with papers outside it; it does not treat the rows as independent samples or estimate how many models a study requires.

Table 11: Model and family counts by research contribution. Papers can belong to several groups. Brackets show interquartile ranges, calculated with interpolation. This gives 4.25 for the training group's upper quartile; the proposed standard instead selects an observed integer and gives 5 (Appendix L). Model counts include names with unspecified releases. CLES is the probability that a paper in the group has more model identifiers than a paper outside it, counting ties as one half.

Contribution stratum	Papers	Identifiers, median [IQR]	Families, median [IQR]	Model-count CLES
New model or method	8,231	3 [1–5]	2 [1–3]	0.53
Training or adaptation	3,816	3 [1–4.25]	2 [1–3]	0.51
Safety or alignment	3,198	3 [1–5]	2 [1–3]	0.54
Benchmark or evaluation	3,005	3 [2–6]	2 [1–4]	0.60
Systems or efficiency	1,896	2 [1–4]	2 [1–3]	0.48
Agents or interaction	1,789	3 [1–5]	2 [1–3]	0.53
Domain application	1,744	3 [1–5]	2 [1–4]	0.53
Other or unclear	760	2 [1–4]	2 [1–3]	0.43

F.2 Does family coverage extend to every task?

A paper can evaluate several model families overall but use only one of them on a particular task family. Here, a *model family* is a related model series, such as GPT or Qwen. A *task family* is a broad evaluation category, such as mathematics or code (Appendix D). We examine the 4,008 papers that evaluate at least two model families and two task families. Together, they contribute 10,809 paper–task-family combinations: each task family counts once within a paper and is counted again when evaluated in another paper. Single-task papers are excluded because their overall and task-specific model-family counts are necessarily the same. Models assigned to the same task family were not necessarily tested on the same dataset or under the same conditions.

For each paper, we calculate the percentage of its task families evaluated with one, two, or at least three model families. We then average these percentages among papers with two, three, four, or at least five model families overall. Every paper receives equal weight, so papers evaluating more task families do not dominate the average. Table 12 reports the results, also shown in Figure 1B. Its first column counts model families across the whole paper; its final three columns describe how many model families are evaluated within each task family.

Table 12: Model-family coverage within task families. Rows group papers by their total number of model families. The final three columns give the average percentage of a paper's task families evaluated with the stated number of model families. These are percentages of task families within papers, not percentages of papers. Totals may differ slightly from 100% because of rounding.

			Average share of task families		
Model families in the whole paper	Number of papers	Paper–task-family combinations	Using 1 model family	Using 2 model families	Using 3+ model families
2	1,477	3,819	33.0%	67.0%	0.0%
3	1,122	2,966	21.7%	19.7%	58.6%
4	628	1,744	17.8%	13.0%	69.2%
5+	781	2,280	15.7%	8.1%	76.2%

Among the 2,531 papers with multiple task families and at least three model families overall, 1,042 (41.2%) evaluate at least one task family using only one model family. This percentage counts papers, whereas Table 12 averages the percentages of task families within each paper. Even among papers with at least five model families overall, an average of 15.7% of task families are evaluated with only one model family.

An alternative calculation gives every paper–task-family combination equal weight. The percentages evaluated with one model family are then 35.3%, 23.4%, 20.0%, and 17.6% across the four groups.

The finding remains: evaluating several model families in a paper does not ensure that several are tested on each task family. Authors should therefore report which models support each main task, alongside the paper-wide total. Supporting experiments may reasonably use fewer models. Our data do not distinguish these experiments from the tasks central to a paper's claim.

F.3 Do the proposed model counts change during the year?

We divide papers into four groups by first submission: September 15–December 14, 2025 (Q1); December 15, 2025–March 14, 2026 (Q2); March 15–June 14, 2026 (Q3); and June 15–September 14, 2026 (Q4). These groups contain 3,043, 2,571, 3,765, and 2,049 papers. Each paper belongs to one period. We recalculate the 75th-percentile counts (R_{75}) for each period, both overall and within the seven contribution groups. We use the same rule for selecting an observed count as in the full-year analysis.

Table 13 shows that family counts are more consistent over time than model counts. Across the four periods, the overall family and task thresholds remain three and two, while the model threshold changes from five to four. Of the 28 contribution-by-period comparisons, 26 retain the full-year family threshold, 19 retain the model threshold, and 20 retain the task threshold. Every difference is one count. These comparisons are not independent because contribution groups overlap.

Table 13: 75th-percentile counts for the full year and each three-month submission period. Each entry gives models/families/tasks.

Group	Full year	Q1	Q2	Q3	Q4
All papers	5/3/2	5/3/2	4/3/2	4/3/2	4/3/2
New model or method	5/3/2	5/3/3	4/3/2	4/3/2	5/3/2
Benchmark or evaluation	6/4/2	6/4/3	6/4/2	6/4/2	6/4/2
Safety or alignment	5/3/2	6/4/3	5/3/2	5/3/2	4/3/2
Domain application	5/4/3	6/4/3	5/4/3	5/4/3	5/4/2
Training or adaptation	5/3/2	5/3/3	4/3/2	4/3/2	4/3/2
Systems or efficiency	4/3/3	5/3/3	4/3/3	4/3/2	4/3/2
Agents or interaction	5/3/2	5/3/3	5/4/2	5/3/2	5/3/2

Benchmark papers retain thresholds of six models and four families throughout the year. Safety papers vary from four to six models and three to four families. The constant overall task threshold also hides changes within contribution groups: benchmark and training papers have a threshold of three tasks in Q1 and two thereafter. A stable overall count therefore does not mean that every group is unchanged.

Table 14 reports the number of papers and 95% intervals from 2,000 bootstrap samples. Each sample draws complete paper records with replacement, keeping that paper's model, family, and task counts together. When many papers have the same count, both interval limits can be identical. Such an interval does not establish error-free extraction.

These comparisons use the latest recovered versions of papers first submitted in each period. They do not reconstruct the evaluations available at four past dates. Earlier papers have had more time for revision, and the mix of topics and contributions can change. The analysis shows how the proposed counts vary with the submission period under the same extraction rules. Family counts are relatively consistent, while model counts need closer attention when the standards are updated.

Table 14: 75th-percentile counts by submission period, with 95% bootstrap intervals. The sample size is the number of papers; contribution groups overlap. Intervals describe variation when the extracted paper records are resampled. They do not measure extraction error or predict future papers.

Group	Cohort	n	Models	Families	Tasks
All papers	Q1	3043	5 [5, 5]	3 [3, 3]	2 [2, 3]
	Q2	2571	4 [4, 5]	3 [3, 3]	2 [2, 2]
	Q3	3765	4 [4, 4]	3 [3, 3]	2 [2, 2]
	Q4	2049	4 [4, 4]	3 [3, 3]	2 [2, 2]
New model or method	Q1	2259	5 [5, 6]	3 [3, 3]	3 [2, 3]
	Q2	1894	4 [4, 5]	3 [3, 3]	2 [2, 2]
	Q3	2718	4 [4, 5]	3 [3, 3]	2 [2, 2]
	Q4	1360	5 [4, 5]	3 [3, 3]	2 [2, 2]
Benchmark or evaluation	Q1	781	6 [6, 7]	4 [4, 4]	3 [2, 3]
	Q2	641	6 [5, 6]	4 [4, 4]	2 [2, 2]
	Q3	1027	6 [5, 6]	4 [4, 4]	2 [2, 2]
	Q4	556	6 [5, 6]	4 [4, 4]	2 [2, 3]
Safety or alignment	Q1	850	6 [5, 6]	4 [3, 4]	3 [2, 3]
	Q2	766	5 [4, 5]	3 [3, 4]	2 [2, 2]
	Q3	1048	5 [4, 5]	3 [3, 3]	2 [2, 2]
	Q4	534	4 [4, 5]	3 [3, 4]	2 [2, 3]
Domain application	Q1	433	6 [5, 6]	4 [3, 4]	3 [3, 3]
	Q2	444	5 [4, 6]	4 [3, 4]	3 [2, 3]
	Q3	547	5 [4, 5]	4 [3, 4]	3 [2, 3]
	Q4	320	5 [4, 5]	4 [3, 4]	2 [2, 3]
Training or adaptation	Q1	1124	5 [5, 6]	3 [3, 3]	3 [3, 3]
	Q2	885	4 [4, 5]	3 [3, 3]	2 [2, 3]
	Q3	1203	4 [4, 4]	3 [3, 3]	2 [2, 2]
	Q4	604	4 [4, 4]	3 [3, 3]	2 [2, 2]
Systems or efficiency	Q1	445	5 [4, 6]	3 [3, 3]	3 [3, 3]
	Q2	451	4 [4, 4]	3 [2, 3]	3 [2, 3]
	Q3	671	4 [4, 4]	3 [3, 3]	2 [2, 3]
	Q4	329	4 [4, 5]	3 [3, 3]	2 [2, 3]
Agents or interaction	Q1	415	5 [5, 6]	3 [3, 4]	3 [2, 3]
	Q2	370	5 [5, 6]	4 [3, 4]	2 [2, 3]
	Q3	641	5 [4, 5]	3 [3, 4]	2 [2, 2]
	Q4	363	5 [4, 5]	3 [3, 3]	2 [2, 2]

G MODEL RELEASE DATES AND AGE AT PAPER SUBMISSION

G.1 MATCHING MODELS TO DATED RELEASES

The release registry draws on provider announcements and model repositories. It records name patterns, family, product series, release order, public release date, access type, and documented parameter counts. More specific names are matched first: GPT-4o-mini, for example, is checked before GPT-4o. A name that does not identify a release is left unmatched.

We define *release age* as the time from the model's public release to the paper's first submission. We also identify whether it was the latest registered release in its product series at submission, and count how many releases had appeared after it within that series. These measures describe different aspects of timing: elapsed days and intervening releases. Neither establishes which model is more capable.

Of 35,790 paper–model records with a release-specific name, 32,233 (90.1%) match the registry. In 178 matched records, the model was released after the paper's first submission. Later paper revisions or ambiguous names and dates can account for these cases. We retain them in the data but exclude them from age calculations, leaving 32,055 records with nonnegative ages. We do not assign ages to missing dates. The results describe the extracted models relative to first submission; they do not establish when the experiments were run.

G.2 RELEASE AGE, NEWER RELEASES, AND USE OVER TIME

Figure 5 separates model age, position within a release series, and monthly family use. The median matched model age is 400 days (13.1 months). Monthly percentages use papers' first submission dates and include family names with unspecified releases. The first and last calendar months cover only part of the month. Changes in family use may reflect changes in paper topics as well as adoption of models.

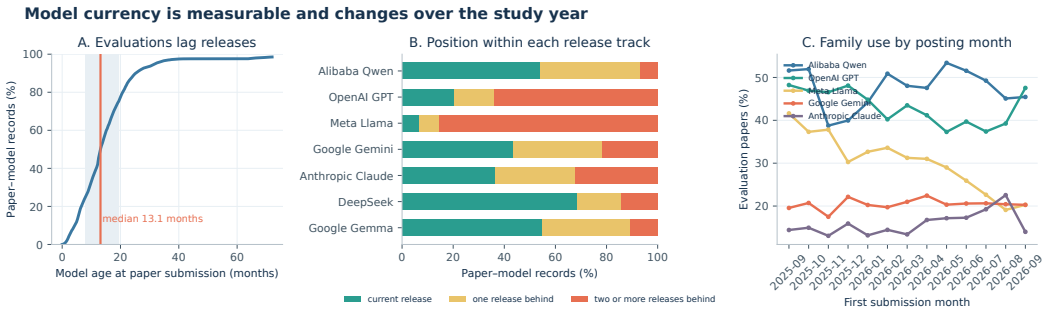


Figure 5: Model timing and use. Panel A shows the proportion of matched records at or below each release age. Panel B shows whether a model was the latest release in its series or how far it was behind. Panel C shows the percentage of evaluation papers using each of the five most frequent families by month. A and B exclude negative ages. C also includes unspecified releases; its family percentages can overlap.

Figure 6 shows differences between families that the overall median hides. The median age is 186 days for Claude and 592 days for Llama. Each family distribution combines its different releases. The results show why researchers should identify the particular release when discussing recency. They do not imply that one family is better or that every study needs a newly released model.

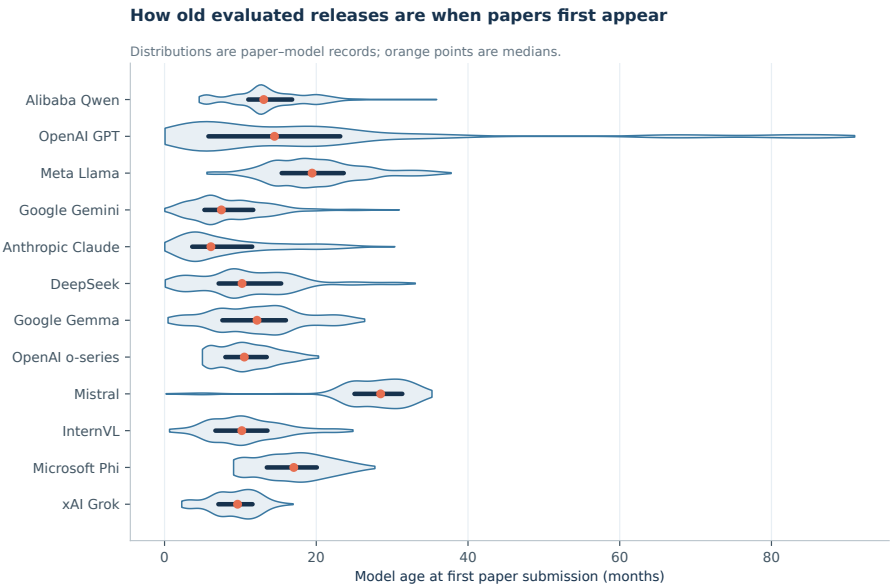


Figure 6: Model age at first paper submission for the twelve most frequent families with matched release dates. Each observation is a paper–model record. Orange points mark medians; dark segments show the middle 50% of ages.

G.3 Larger panels, newer models, and access types

In Figure 7, the median age of a paper's newest matched model falls from 14.5 months for one-family papers to 5.3 months for papers with at least five families. The four groups contain 3,241, 2,525, 2,902, and 1,145 papers with a release match. Papers with more families are also more likely to include a recent model and both open-weight and hosted models. A larger panel has more chances to include any of these properties, so the comparison does not isolate an effect of panel size.

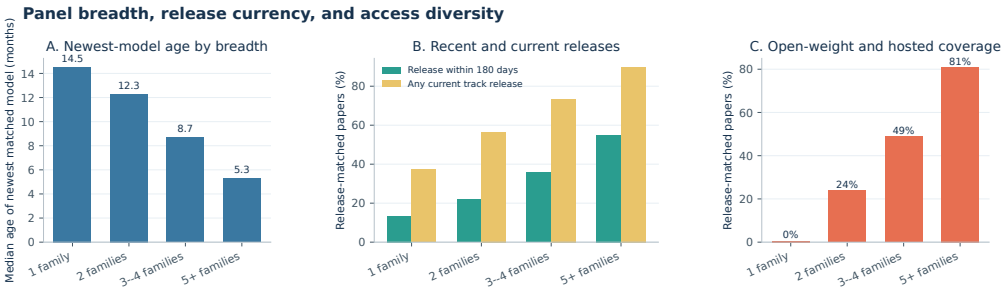


Figure 7: Model age and access types by number of evaluated families. The newest-model age is the smallest matched age in each paper. Percentages for recent or current releases and for access types use papers with at least one release match. These are observed associations; the comparison does not adjust for the number of opportunities a larger panel has to include each property.

This figure uses 180 days as a rounded window for recent releases. The proposed overall age standard is 170 days, derived from the observed age distribution. Task-specific standards consider only models evaluated on that task (Appendix L). These calculations use different thresholds and sets of records, so their percentages should not be treated as interchangeable.

H WHAT REPORTED PERFORMANCE CAN TELL US ACROSS PAPERS

A model's position in a results table depends on which other models the authors tested. Papers also use different tasks, datasets, and evaluation settings. This section examines how those differences affect summaries of performance across papers. We compare ranking methods, test predictions on papers not used to fit the comparison model, and examine how estimates change when papers are resampled. These analyses are separate from the proposed standards for model counts and recency. Their purpose is to assess what the published results can support, rather than produce a leaderboard as if all models had undergone the same evaluation.

H.1 EXTRACTED SCORES AND WHAT WE SUMMARIZE

The score records contain the original text of each score, its numerical value, whether higher or lower is better, its table and column, the caption, task labels, and model name. A *comparison block* is a column containing at least two models, an assigned task, and a known direction of improvement. The parser removes percent signs and reads the value before a $\pm$ term. It does not carry the reported uncertainty into the analysis. When the same normalized model name appears more than once in a column, it retains the first extracted row. This rule is reproducible, but it neither averages different configurations nor selects the best result. Differences between configurations remain unresolved.

We rank models within each block, assigning average ranks to ties. For model i in a block of m_b models, let $r_{ib} = 1$ denote the best rank. Its percentile is $(m_b - r_{ib})/(m_b - 1)$: one for the best untied result and zero for the worst. We then take the median over blocks for the same paper, model, and task. This reduces 50,976 rank observations to 6,277 paper-level summaries. For a family summary, we combine its members' ranks within each paper and task and take the median. We do not average raw scores from benchmarks with different scales.

H.2 WHY THE CHOICE OF OTHER MODELS MATTERS

When higher scores indicate better performance, the percentile can also be written as

$$q_{ib} = \frac{m_b - r_{ib}}{m_b - 1} = \frac{1}{m_b - 1} \sum_{j \neq i} \left[\mathbb{1}(s_{ib} > s_{jb}) + \tfrac{1}{2} \mathbb{1}(s_{ib} = s_{jb}) \right]. \tag{1}$$

This equation says that the percentile is the proportion of other models beaten, with ties counting as one half. The equality holds for all 50,976 rank observations. Simply replacing a percentile with a win rate therefore does not remove dependence on the models compared, although changing how results are averaged can change their weights. For example, beating the only other model gives a percentile of one. Adding four stronger models reduces it to 0.2, even though the original comparison is unchanged.

H.3 ADJUSTING FOR WHICH MODELS WERE COMPARED

Chatbot Arena uses the Bradley–Terry model to estimate relative model strength from paired human judgments (Chiang et al., 2024). We use the same statistical form to examine the ordering of published scores. Here, a "win" means a higher reported point estimate after accounting for metric direction. It is not a human preference judgment or evidence that the difference is statistically significant.

The 12,362 comparison blocks from 1,441 papers provide 131,055 unordered model pairs. We assign one for a win, zero for a loss, and one half for a tie. Repeated comparisons of the same pair are averaged within each paper and task, leaving 14,219 paper–task–pair records. Within each task, every paper receives total weight one, shared equally across its observed model pairs. This prevents papers with many comparisons from dominating the fit.

For task t, we estimate a strength coefficient β_{it} for each model, with $\sum_i \beta_{it} = 0$, by minimizing

$$\sum_p \frac{1}{|E_{pt}|} \sum_{(i,j) \in E_{pt}} \{\log[1 + \exp(\beta_{it} - \beta_{jt})] - y_{pijt}(\beta_{it} - \beta_{jt})\} + \frac{\lambda}{2} \sum_i \beta_{it}^2, \tag{2}$$

where E_{pt} is the set of model pairs in paper p on task t, and y_{pijt} is the average win score for model i against model j. The first term measures agreement with the observed comparisons. The final term

limits extreme strength estimates; λ controls this penalty. We use $\lambda = 0.1$ and check 0.01 and 1. The calculation accepts fractional average outcomes and does not treat score columns as independent preference trials.

We fit each task separately. Within a task, models can be connected directly by a comparison or indirectly through shared competitors. We fit disconnected groups separately because their relative strengths cannot be inferred from observed comparisons between them.

Before fitting, we repeatedly remove models compared with the remaining models in fewer than five papers. This rule uses comparison availability, not win outcomes. It leaves 224 model–task combinations and 1,977 paper–task–pair records from 625 papers. Eleven task families retain a connected group of at least three models; science does not. In each retained group, every model can reach every other through a sequence of recorded wins, the condition called strong connectivity. However, in the unrestricted task networks, 88.6–93.7% of directly compared pairs appear in only one paper. A connected network therefore need not provide repeated evidence for its individual comparisons. Five papers is a threshold for this analysis, not a validated minimum for reliable estimation.

H.4 Comparing methods on the same evidence

To compare rankings fairly, we recalculate percentiles using the same models and blocks retained for each Bradley–Terry fit. We also express each fitted strength as its average predicted win probability against the other models in its connected group, giving every competitor equal weight. This provides a common set of competitors under the model's assumptions. Predictions for pairs that were never compared depend on those assumptions. We do not compare fitted scores across different tasks or disconnected groups.

We estimate uncertainty by drawing 1,000 bootstrap samples of papers and refitting at $\lambda = 0.1$, keeping the set of models fixed. All pairs from a sampled paper receive the same sampling weight. Between 1.1% and 6.9% of these samples disconnect the comparison network. Some relative positions then depend on the penalty rather than observed links. Across the 224 model–task combinations, the median width of the 95% rank interval is 15 positions.

Table 15 shows how the two methods agree when they use the same evidence. Correlations range from 0.50 to 0.95, with a median of 0.76. Of 2,005 pairwise orderings that are not tied under percentiles, 387 reverse under Bradley–Terry adjustment (19.3%). These reversals describe related changes in rankings; they are not independent hypothesis tests. Comparing adjusted scores with percentiles from the unrestricted data gives correlations of 0.34–0.87, but that comparison changes both the method and the included evidence.

Table 15: Changes in model ordering after adjusting for competitors. Nodes are retained model identifiers, and papers are those supporting the comparison network. The correlation ρ compares percentiles and Bradley–Terry scores calculated on the same evidence. Reversals exclude pairs tied under percentiles. CI width is the median width of the 95% bootstrap rank interval, in positions. Science has no group meeting the inclusion rule.

Task	Models	Papers	ρ	Reversals (%)	CI width
Agents and robotics	8	26	0.95	4.0	4.5
Code	18	56	0.77	16.2	11.5
Language	29	92	0.69	21.6	19.0
Mathematics	12	46	0.90	10.0	8.0
Medicine	32	78	0.69	23.3	20.0
Multimodal	22	63	0.80	14.0	12.0
Professional applications	21	52	0.76	17.6	14.0
General reasoning	8	25	0.75	18.5	5.0
Retrieval and context	14	42	0.78	11.6	10.0
Science	0	—	—	—	—
Models and systems	28	89	0.62	17.0	19.0
Safety and society	32	91	0.50	23.6	25.0

Several additional checks show which choices affect the ordering. Changing the penalty gives correlations of 0.85–1.00 with the main fit. Giving each observed pair equal weight, rather than each paper, has a larger effect: correlations are 0.32–0.95. Requiring every pair to appear in at least three papers splits ten of the eleven networks into disconnected groups. Removing the most frequently used comparison model gives correlations of 0.69–0.94; removing names without a specified release gives 0.70–1.00.

We also fit families rather than individual models. These fits average comparisons between different families within blocks and papers and exclude comparisons within a family. Their correlations with the original family percentiles range from 0.27 to 1.00. Because each family combines several releases, the fitted value does not describe one system's ability.

Lowering the requirement from five papers per model to three retains 606 model–task combinations across all twelve tasks. Correlations between methods on the same evidence are then 0.61–0.80, with a median of 0.74. Raising it to ten leaves no connected group with at least three models after repeated filtering. Thus, frequent model use does not guarantee enough repeated comparisons to establish an ordering. Which comparisons are retained and how they are weighted both matter.

H.5 Can the fitted ordering predict results in other papers?

We use five-fold evaluation: the papers are divided into five groups, each is held out in turn, and the comparison model is fitted using the other four. Within each task, all observations from a paper stay together. The set of eligible models is fixed using comparison availability, but fitted strengths use only the papers assigned to fitting in that fold.

For comparison, we construct a predictor from median percentiles in those same fitting papers, using the matched model sets. Percentiles are restricted to $[0.05, 0.95]$ before conversion to log odds. A nonnegative scaling coefficient for pairwise differences is estimated on the fitting data. Both models in a tested pair must belong to the same connected group in the fitting data; all 1,977 retained records meet this condition.

We compare this percentile predictor and the three Bradley–Terry penalties with a neutral prediction of 0.5 for every pair. The measure is cross-entropy, which is lower when predicted probabilities better match the observed outcomes. We average it first within each paper and task, then across tasks within each paper, and finally across the 625 papers. Two thousand paper-bootstrap samples give intervals for prediction losses and their paired differences. These intervals resample the existing held-out predictions; they do not repeat fitting, fold assignment, or model selection.

Table 16 shows that the main Bradley–Terry fit predicts worse than the neutral model. The strongest penalty has the lowest estimated loss, but its difference from neutral includes zero. The percentile predictor also has no clear advantage. These comparisons therefore do not establish a reliable ordering that transfers across the recovered papers. The finding concerns heterogeneous published scores, not preference judgments collected under a common evaluation protocol.

Table 16: Prediction of score orderings in held-out papers. Lower cross-entropy is better. The final column gives the loss difference from a neutral prediction of 0.5, with a 95% paper-bootstrap interval. Fractional outcomes summarize reported score comparisons rather than individual preference votes.

Method	Cross-entropy	Δ from neutral [95% interval]
Matched percentiles	0.706	+0.013 [-0.002, +0.029]
Bradley–Terry ($\lambda = 0.01$)	0.749	+0.056 [+0.032, +0.082]
Bradley–Terry ($\lambda = 0.1$)	0.721	+0.028 [+0.010, +0.047]
Bradley–Terry ($\lambda = 1$)	0.687	-0.006 [-0.015, +0.003]
Neutral (0.5)	0.693	+0.000 [+0.000, +0.000]

H.6 Shared comparison models and response length

Another approach would compare every model against the same baselines. For each task, we select the two identifiers most frequently used in direct comparisons, breaking frequency ties alphabetically. A model must have at least five direct-comparison papers against each baseline. We average its win

scores across papers for each baseline, then give the two baselines equal weight. Only one model–task combination meets this requirement, so the rule does not support a comparison among several models.

Length-Controlled AlpacaEval addresses a different source of bias: evaluators may prefer longer responses. It estimates preferences at equal response length using paired responses, evaluator judgments, and length differences (Dubois et al., 2024). Our extracted aggregate scores do not include those data. Parameter counts, context-window sizes, and token budgets cannot substitute for the lengths of the actual responses.

Table 17 summarizes what the approaches measure. We can examine competitor adjustment with the available data, subject to the comparison model's assumptions. We cannot estimate an adjustment for response length.

Table 17: What the comparison methods measure and what data they require. Accounting for competitor choice and accounting for response length address different problems.

Method	What it measures	Limitation in this study
Within-panel percentile	Relative position among the models selected by a paper	Changing competitors can change the percentile without changing an original pairwise result.
Bradley–Terry	Probability of beating another model under a shared-strength model within a task and connected group	Adjustment depends on assumptions and cannot remove unobserved selection or differences in evaluation settings.
Common-opponent win rate	Average result against the same set of comparison models	Too few models have repeated direct comparisons with both selected baselines.
Length-controlled win rate	Evaluator preference at equal response length	Requires response lengths and judgments absent from the extracted data; it does not itself standardize the models compared.

H.7 Performance distributions in the original papers

Figures 8 and 9 show the ten most frequently evaluated release-specific identifiers across all twelve task families. The amount of evidence differs greatly: GPT-4o has 34 papers in language but five in mathematics, and some model–task combinations have no papers. A box is drawn when there are at least three papers; this is a display rule, not a claim of adequate evidence.

The spread of each box reflects differences across the source papers, including their comparison models and evaluation settings. It is not an uncertainty interval from one shared evaluation protocol. The figures describe the results available in the literature; the adjusted and predictive comparisons above explain why they should not be read as a general model ranking.

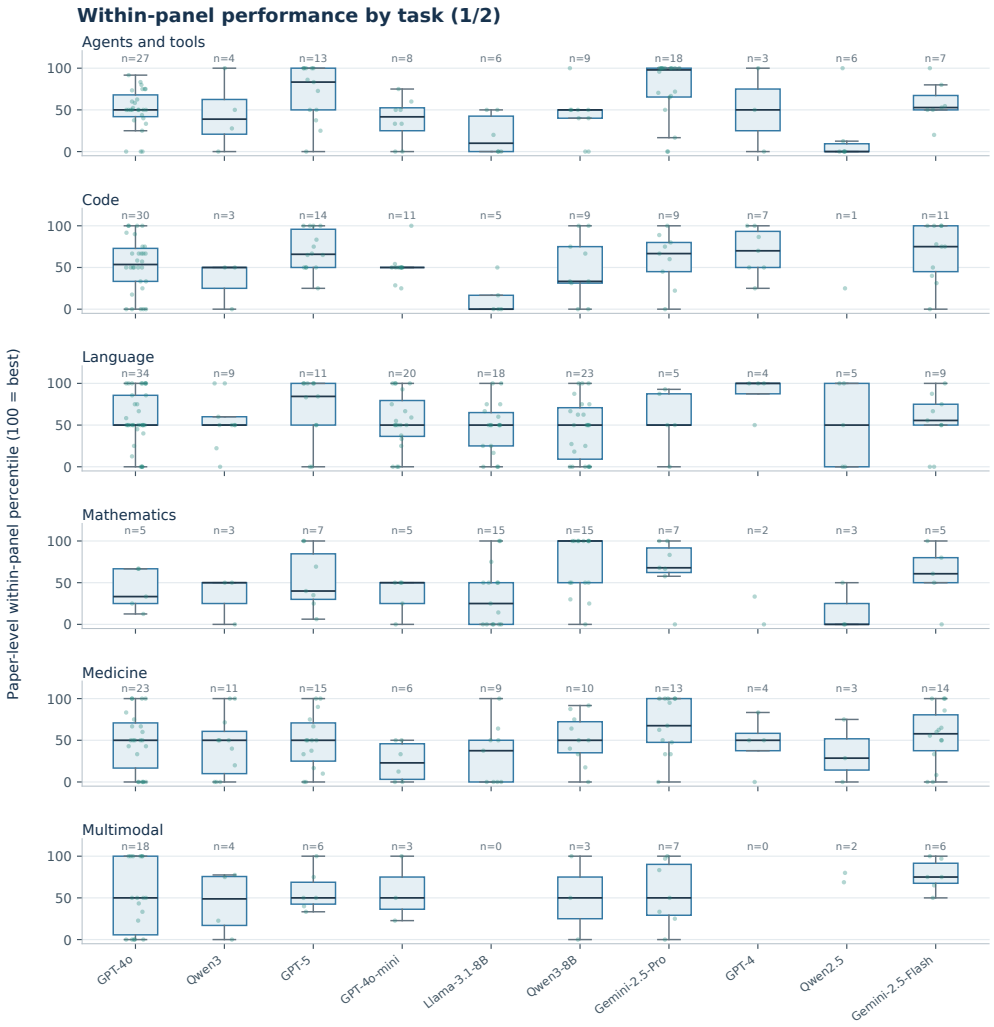


Figure 8: Percentile positions in source papers for the ten most frequently evaluated release-specific identifiers, covering task families 1–6. Each observation is a paper-level median on a 0–100 scale. Boxes require at least three papers and show the median and interquartile range (IQR). Whiskers extend to observations within 1.5 IQRs. Counts identify sparse and empty groups. Differences may reflect the competitors selected as well as model performance.

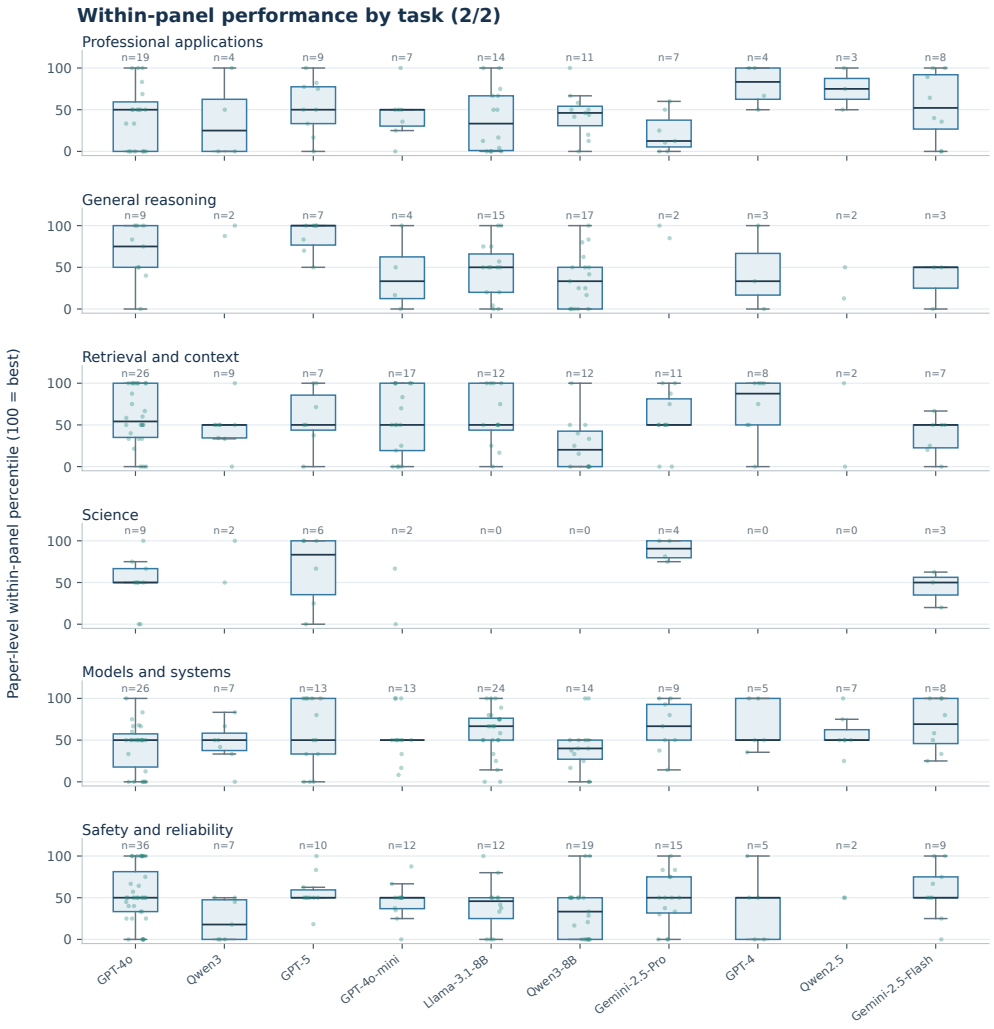


Figure 9: Percentile positions for the same ten model identifiers, covering task families 7–12. Display rules match Figure 8. All task families are shown, including those with little evidence.

H.8 How many papers give a stable median?

The initial analysis uses the same 34 family–task combinations at every sample size, each with at least 40 papers. For each combination, we draw two samples without replacement and with no papers shared between them. We repeat this 2,000 times and calculate the absolute difference between their median percentiles. Although the samples do not overlap, their selection is dependent: choosing a paper for the first sample removes it from the second sample's available pool. We combine the resulting differences with the same number of draws from every family–task combination.

Table 18 reports the disagreement between these paired medians. We set a tolerance of ten percentile points and ask when at least 75% of the differences are within that tolerance. Fifteen papers is the first tested sample size to meet it. This is a chosen criterion for stability of the median, not a power calculation or a minimum number of papers for every study.

Table 18: Differences between median percentiles from two non-overlapping samples of papers. Differences are absolute and measured in percentile points. Each family–task combination contributes equally to the summary.

n papers	Median	75th pct.	90th pct.	Within 10 pp
3	16.7	33.3	50.0	38.7%
5	11.1	25.0	41.7	49.4%
10	5.2	13.3	25.0	67.0%
15	0.0	10.0	16.7	77.1%
20	0.0	6.2	14.6	83.6%

Figure 10 compares this design with two independently drawn samples with replacement, using 20,000 repetitions per combination. With replacement, the 75th-percentile disagreement is 11.1 points at 15 draws and 8.3 at 20 draws. Twenty is therefore the first tested size that meets the tolerance. The required number depends on the sampling design; where a single planning value is needed, the with-replacement value of 20 papers is the conservative choice. Neither design removes dependence on the original comparison models: stable medians can still reflect a stable pattern of competitor selection.

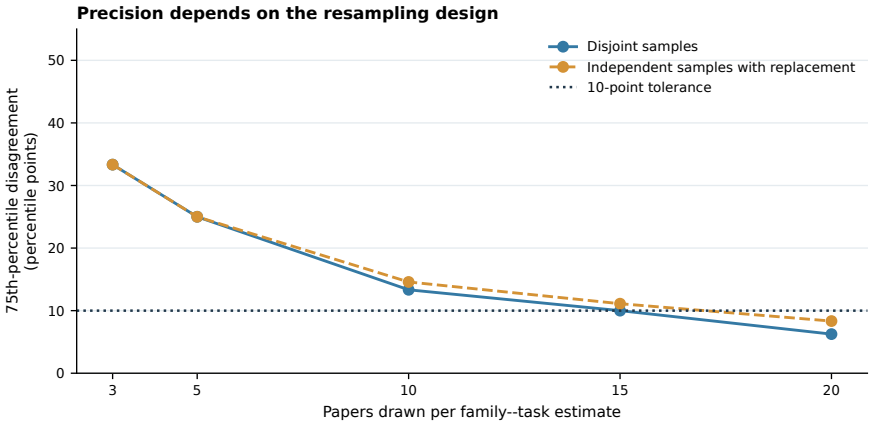


Figure 10: Disagreement between sample medians for the same 34 family–task combinations. The curves compare non-overlapping samples drawn without replacement with independent samples drawn with replacement. The dotted line marks the chosen tolerance of ten percentile points.

Effects of small comparison sets. Of the 12,362 blocks, 4,583 (37.1%) contain only two models. Across 4,365 paper–family–task summaries, 31.5% have a median of exactly 50 on the 0–100 scale. Requiring at least three models per block lowers that share to 26.9%. Small comparison sets allow few rank values, so repeated median values may partly reflect this limited resolution and ties.

Table 19 examines whether removing two-model blocks changes stability. It uses the 23 family–task combinations that still have at least 40 papers after the restriction. With samples of 15, the restricted blocks give 75th-percentile disagreement of 8.3 points without replacement and 10.0 with replacement. Results are similar when all blocks are used from those same retained papers. Removing two-model blocks alone therefore does not remove the stability pattern. This restricted comparison does not replace the result from all 34 combinations. Requiring five models per block leaves no combination with 40 papers; the largest has 35.

Table 19: Effect of excluding two-model blocks, using the same 23 family–task combinations and 20,000 draws per combination. Entries are the 75th percentiles of absolute differences between sample medians, in percentile points. *Same papers* uses all blocks but restricts papers to those retained by the three-model rule.

Evidence	$n = 15$, disjoint	$n = 15$, replacement	$n = 20$, disjoint	$n = 20$, replacement
All blocks, same cells	5.7	8.3	3.7	5.0
All blocks, same papers	8.3	10.0	5.2	7.1
Blocks with ≥ 3 models	8.3	10.0	5.4	7.3

Implications and reproducibility. A summary across papers should report how many papers support it, which models were directly compared, how results were weighted, and the uncertainty. A stable median or a connected comparison network alone does not establish a model's general ability. The code uses seed 20260921 for network fits and fold assignment and 20260922 for prediction-loss resampling. It generates pair records, network checks, fitted scores, bootstrap draws, held-out predictions, and shared-baseline checks. Implementation tests recovered simulated strengths, reversed coefficients when wins were reversed, and detected disconnected groups. These tests check the calculations, not the accuracy of extracted scores or the comparability of unreported settings.

I Comparison with papers in the ICLR community

I.1 Included papers and adjustment for group differences

The ICLR 2026 proceedings contain 5,351 papers. Matching normalized titles links 2,756 of them to the arXiv study population: 2,739 exact title matches and 17 high-confidence approximate matches. None of the approximate matches belongs to the eligible evaluation set. All 486 ICLR papers in this analysis therefore have exact normalized-title matches. We compare them with the remaining 10,942 eligible arXiv papers. That group is not a set of rejected ICLR submissions.

The two groups differ in timing and subject matter. We account for measured differences by weighting papers so that the comparison emphasizes those with similar characteristics. A ridge-logistic model estimates each paper's probability of belonging to the ICLR group from first-submission month, primary arXiv category, task labels, and contribution labels. Categories with fewer than 40 papers are combined as "other." The model has 50 columns: an intercept, 12 month indicators, 17 category indicators, 12 task indicators, and eight contribution indicators. The first sorted month and category are the omitted comparison levels.

The fitted model minimizes the summed negative log likelihood plus $\lambda \sum_{j>0} \beta_j^2/2$, with $\lambda = 0.5$. The penalty limits extreme coefficients and does not apply to the intercept. Estimated probabilities are restricted to $[.005, .995]$. If $\hat{e}(X)$ is the estimated probability of ICLR membership, an ICLR paper receives weight $1 - \hat{e}(X)$ and another arXiv paper receives weight $\hat{e}(X)$. This procedure, called overlap weighting, gives greater influence to papers whose measured characteristics occur in both groups. The result describes that weighted comparison, rather than a difference between all ICLR and all arXiv papers.

We check how well weighting aligns the measured characteristics. The largest difference in their weighted proportions is 0.0168. The largest standardized difference, calculated using the unweighted pooled standard deviation, is 0.067. Effective sample sizes, $(\sum w)^2 / \sum w^2$, are 479 for ICLR and 3,225 for the comparison group; this measure summarizes how much the weights concentrate the available information.

The main uncertainty intervals use 2,000 bootstrap samples drawn separately within the two groups. Each sample refits the probability model, using the same set of category labels. Additional checks vary the penalty from 0.05 to 5. We also fit a second model with 38 columns that omits the twelve task indicators. This matters because making task-label frequencies similar also makes their sum—the task count—similar. The second model allows us to examine task-count differences without directly balancing those labels.

I.2 Differences in model counts, families, and access

Table 20 gives the main weighted comparison. ICLR papers include 0.53 more model identifiers and are 6.2 percentage points more likely to include both open-weight and hosted models. The estimated difference in family count is 0.13, with an interval including zero. Here, a recent model means one released within 180 days before first submission. This rounded comparison window differs from the proposed overall age standard of 170 days. Release-age outcomes require a matched release date.

Table 20: Weighted comparison between ICLR papers and other eligible arXiv papers. Values are means or percentages after overlap weighting. Differences are ICLR minus the comparison group, calculated before rounding. The 95% intervals resample papers within each group and refit the probability model in every bootstrap sample.

Measure	ICLR 2026	Matched arXiv	Adjusted difference [CI]
Model identifiers	4.45	3.92	+0.53 [+0.21, +0.87]
Model families	2.64	2.51	+0.13 [-0.02, +0.29]
A model released within 180 days	36.6%	32.5%	+4.2 pp [-0.1, +8.7]
Both open-weight and hosted models	33.8%	27.6%	+6.2 pp [+2.3, +10.5]
Three families and a recent model	24.6%	20.1%	+4.5 pp [+0.7, +8.6]

When task indicators are omitted, the differences are 0.74 model identifiers, 0.21 families, and 0.21 tasks; the task interval is [0.09, 0.32]. Additional models from already represented families account for 76% of the identifier difference in the main model and 72% in the model without task indicators. Figure 11 compares the two sets of estimates, and Table 21 gives their intervals. Both associate ICLR appearance mainly with more models within families. Differences in family and task counts depend more on what is balanced. The two weighting models also emphasize somewhat different sets of papers.

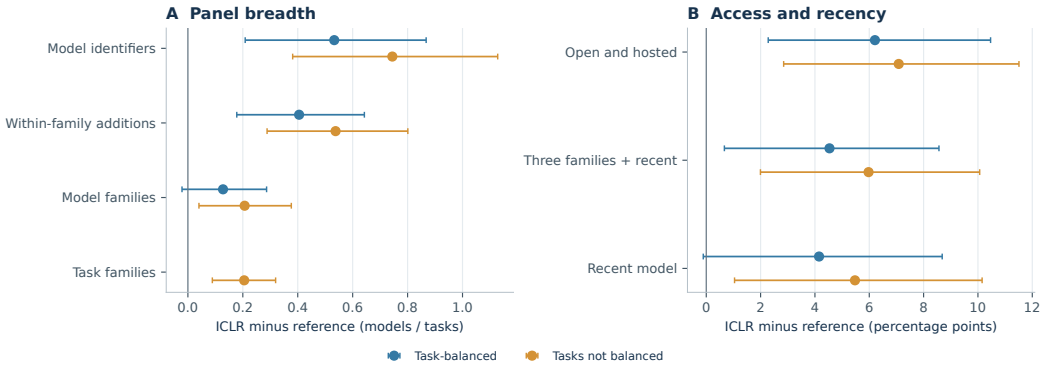


Figure 11: Differences between ICLR and other arXiv papers under two weighting models. Both account for timing, arXiv category, and contribution; blue also balances task labels. Each interval uses 2,000 bootstrap samples that refit the corresponding model. The task-count estimate is omitted from the task-balanced model because balancing task labels directly constrains that count.

Table 21: Comparison with and without balancing task labels. Entries are ICLR minus other arXiv papers, with 95% intervals from 2,000 bootstrap samples that refit the weighting model. Percentage differences are in percentage points. Balancing task indicators directly constrains the task-count difference.

Measure	Task-balanced	Tasks not balanced
Model identifiers	+0.53 [+0.21, +0.87]	+0.74 [+0.38, +1.13]
Model families	+0.13 [-0.02, +0.29]	+0.21 [+0.04, +0.38]
Within-family additions	+0.40 [+0.18, +0.64]	+0.54 [+0.29, +0.80]
Task families	Balanced by construction	+0.21 [+0.09, +0.32]
A model released within 180 days	+4.2 pp [-0.1, +8.7]	+5.5 pp [+1.0, +10.2]
Both open-weight and hosted models	+6.2 pp [+2.3, +10.5]	+7.1 pp [+2.8, +11.5]
Three families and a recent model	+4.5 pp [+0.7, +8.6]	+6.0 pp [+2.0, +10.1]

I.3 Distributions and submission-period checks

Tables 22 and 23 examine whether the higher model count comes only from a few unusually large ICLR panels. The medians are the same, but the trimmed and geometric means are higher for ICLR. Upper percentiles and the proportions of papers above specified model counts also place more of the difference among larger panels. Unlike the main analysis, these additional intervals keep the fitted weights fixed. They help describe the distributions but do not include uncertainty from re-estimating the weights.

Table 22: Unweighted and weighted summaries of the two groups. Trimmed means remove the lowest and highest 10% before averaging. Geometric means are calculated after adding one and then transformed back. Within-family additions equal the model identifier count minus the family count. Table 23 reports the corresponding weighted differences.

Measure	Summary	Raw ICLR / arXiv	Adjusted ICLR / arXiv
Exact identifiers	Mean	4.49 / 3.44	4.45 / 3.92
	Median	3.00 / 3.00	3.00 / 3.00
	10% trimmed mean	3.79 / 2.86	3.74 / 3.33
	Geometric mean	3.46 / 2.73	3.43 / 3.11
	75th percentile	6.00 / 4.00	6.00 / 5.00
	90th percentile	10.00 / 7.00	10.00 / 8.00
Within-family additions	Mean	1.85 / 1.09	1.81 / 1.41
	Median	1.00 / 0.00	1.00 / 1.00
	10% trimmed mean	1.28 / 0.65	1.25 / 0.94
Families per identifier	Mean	73.4% / 81.2%	73.9% / 77.2%
Any repeated family	Share	57.2% / 43.5%	56.6% / 51.3%
At least six identifiers	Share	28.8% / 17.4%	28.5% / 22.8%
At least ten identifiers	Share	10.5% / 4.9%	10.3% / 6.8%
Model families	Mean	2.64 / 2.36	2.64 / 2.51
	Median	2.00 / 2.00	2.00 / 2.00
	75th percentile	4.00 / 3.00	4.00 / 3.00
Task families	Mean	2.14 / 1.77	2.09 / 2.07
	Median	2.00 / 1.00	2.00 / 2.00
	75th percentile	3.00 / 2.00	3.00 / 3.00

Table 24 compares papers within task families before weighting. The median model identifier count is at least as high for ICLR in every task, whereas median family counts show a less consistent difference. The percentage including a recent model is higher for ICLR in all twelve task groups. Its smaller difference after adjustment shows why timing and research topic matter when interpreting the raw comparison.

Table 25 repeats the analysis within shorter first-submission periods for both groups. The estimated model-count difference stays near half a model. Differences in coverage of both access types

Table 23: Weighted ICLR-minus-arXiv differences for Table 22. The 95% bootstrap intervals keep the fitted weights fixed; the main intervals in Table 20 instead refit them. The Benjamini–Hochberg procedure adjusts the bootstrap p values together across the 19 comparisons; the adjusted values are reported as q.

Measure	Summary	Difference [95% CI]	BH q
Exact identifiers	Mean	+0.53 [+0.19, +0.90]	0.032
	Median	+0.00 [+0.00, +1.00]	1.000
	10% trimmed mean	+0.41 [+0.10, +0.75]	0.047
	Geometric mean	+0.32 [+0.07, +0.59]	0.047
	75th percentile	+1.00 [+0.00, +2.00]	0.148
	90th percentile	+2.00 [+0.00, +3.00]	0.070
Within-family additions	Mean	+0.40 [+0.16, +0.66]	0.032
	Median	+0.00 [+0.00, +1.00]	1.000
	10% trimmed mean	+0.31 [+0.11, +0.52]	0.032
Families per identifier	Mean	-3.3 pp [-5.8, -1.0]	0.044
Any repeated family	Share	+5.2 pp [+0.6, +10.1]	0.061
At least six identifiers	Share	+5.7 pp [+1.5, +10.0]	0.043
At least ten identifiers	Share	+3.5 pp [+0.8, +6.3]	0.047
Model families	Mean	+0.13 [-0.04, +0.29]	0.209
	Median	+0.00 [+0.00, +0.00]	1.000
	75th percentile	+1.00 [+0.00, +1.00]	0.484
Task families	Mean	+0.01 [-0.11, +0.13]	1.000
	Median	+0.00 [+0.00, +0.00]	1.000
	75th percentile	+0.00 [-1.00, +0.00]	1.000

Table 24: Task-specific comparisons before adjustment. A paper can appear in more than one task group. Recent-model percentages use papers with at least one matched release date. These rows describe observed practice within each task; they do not replace the weighted comparison.

		Median exact models		Median families		Recent model
Task family	ICLR n	ICLR	arXiv	ICLR	arXiv	ICLR / arXiv
Models and systems	167	3	3	2	2	40.5% / 20.8%
Language and communication	134	4	3	2	2	41.7% / 25.1%
Mathematics and formal reasoning	122	4	3	2	2	43.5% / 26.4%
Safety, reliability, and society	118	5	3	3	3	50.0% / 29.9%
Multimodal perception and generation	100	3	3	3	2	43.7% / 23.1%
Knowledge and general reasoning	89	4	4	2	2	45.8% / 23.8%
Code and software engineering	70	4	3	3	2	51.5% / 35.9%
Professional and social applications	66	4.5	3	2.5	2	63.9% / 34.2%
Agents, tools, and robotics	54	4	3	3	2	53.8% / 37.2%
Retrieval, memory, and long context	50	5	3	3	2	51.0% / 26.7%
Medicine and health	42	4.5	3	3	2	53.8% / 34.1%
Science and discovery	28	4	3	3	2	65.4% / 42.4%

remain positive, while family-count differences remain small. We refit the weighting model for each restricted period, then keep those weights fixed in its bootstrap samples. This tests dependence on the submission window but does not include the full weight-estimation uncertainty used in the main analysis.

Table 25: Comparisons within restricted first-submission periods. A new weighting model is fitted for each period. Bootstrap intervals then hold those weights fixed, unlike the main analysis, which refits them in each sample. Outcomes involving model age require a matched release date.

First-posting cutoff	Measure	ICLR n	Adjusted difference	95% interval
2025-12-31	Exact model identifiers	357	+0.47	[+0.00, +0.90]
2025-12-31	Model families	357	+0.12	[-0.09, +0.33]
2025-12-31	Both open-weight and hosted	357	+6.7 pp	[+1.1, +12.2]
2025-12-31	Three families and a recent model	328	+7.9 pp	[+2.4, +13.6]
2026-03-31	Exact model identifiers	464	+0.53	[+0.15, +0.93]
2026-03-31	Model families	464	+0.11	[-0.07, +0.29]
2026-03-31	Both open-weight and hosted	464	+6.9 pp	[+2.3, +11.6]
2026-03-31	Three families and a recent model	424	+4.9 pp	[+0.8, +9.6]

I.4 What the ICLR comparison supports

The comparison includes arXiv-linked papers and may use versions revised after review. It does not reconstruct the evaluation panels originally seen by reviewers. Title matching, arXiv posting, author resources, and unmeasured research differences can affect the two groups. Review decisions also vary (Cortes & Lawrence, 2021; Beygelzimer et al., 2021). We therefore treat ICLR appearance as a characteristic of the published paper, not a measure of scientific quality or an outcome caused by panel size.

Applying the same extraction rules to both groups does not guarantee equal error across formats or reporting styles. Weighting cannot correct unmeasured extraction error. Within these limits, the comparison describes evaluation practice in the ICLR community: the clearest differences are more variants within families and greater inclusion of both access types, rather than a consistent increase in models, families, and tasks alike.

J How evidence requirements affect the results

We examine whether the proposed counts depend on which evaluation records are included. One analysis broadens the inclusion rules to admit additional records. Another requires stronger evidence for records already included. Both use the same text extraction and model-name dictionary. Agreement between them can show that a result is less sensitive to these choices, but it does not independently establish that the original classifications were correct.

J.1 Allowing additional evidence

Table 26 compares three rules, each including all records accepted by the preceding one. The primary rule requires the implemented evidence of author execution and task assignment. The inclusive rule also accepts prose connecting author use with a nearby result. The permissive rule accepts named rows in results tables when execution is inferred from their position, without a separate statement. Across these rules, the number of papers increases from 11,428 to 13,748 and the number of model–task records from 65,863 to 97,273.

These differences show how inclusion depends on the required evidence. The primary rule is the most restrictive of these three. The label “strict” describes that rule; it does not mean every included evaluation has been independently verified.

Table 26: Counts under progressively broader inclusion rules. All three aim to exclude model mentions without evaluation, copied results, and models used only in supporting roles.

Evidence rule	Papers	Paper–model–task records
Strict author-tested rule	11,428	65,863
Inclusive local-result context	12,970	77,809
Permissive named result rows	13,748	97,273

J.2 Requiring stronger evidence for included records

For each included paper–model–task record, we examine all evidence supporting it. We then apply four filters: a readable numerical table result; no recorded row citation unless rerunning is explicit; a task label recoverable from nearby text under the original rules; or all three requirements together. A record remains if at least one supporting entry passes the filter.

The nearby-text filter excludes titles and abstracts, but does not reconstruct the exact route used for the original task assignment. Citation detection is less complete in PDF and HTML than in LaTeX. Requiring numbers can also exclude legitimate prose or qualitative evaluations. These filters therefore select different kinds of evidence, rather than supplying verified labels.

Table 27 distinguishes two changes: removing records within a paper and removing papers that have no records left. Each row shows results for papers retaining at least one record, together with those papers' original 75th-percentile model count. For example, the numerical-table filter retains 3,512 papers. Their original 75th-percentile count was six models; after filtering, it is four.

Table 27: Results after requiring stronger evidence. Medians and 75th-percentile counts (R_{75}) use papers with at least one retained record. The final column gives the original model R_{75} for those same papers, helping distinguish the effect of removing evidence from the effect of selecting papers. M/F/T means models/families/tasks.

Evidence	Relations	Papers	Median M/F/T	R_{75} M/F/T	Prior model R_{75}
Primary	65,863	11,428	3/2/1	5/3/2	5
Numeric table	18,216	3,512	2/2/1	4/3/2	6
Citation screen	64,510	11,361	3/2/1	4/3/2	5
Local task	29,498	6,809	2/2/1	4/3/2	5
All three screens	8,217	2,119	2/2/1	4/3/2	6

Under all four filters, the 75th-percentile family count remains three and the task count remains two. The model count falls from five in the full analysis to four. The top five families account for 67.0–69.1% of paper–family records. Thus, concentration in a few families and the family-count threshold persist under these restrictions. The model-count threshold is more sensitive to which evidence is retained.

Papers with no records left are excluded from the filtered summaries, not counted as evaluating zero models. The supplement provides record-level flags for repeating these checks. Their purpose is to identify results that depend on evidence choices, not to detect every extraction error or establish that one supporting record is sufficient for validity.

K How evaluations were extracted and checked

K.1 Extraction procedure

The pipeline uses name dictionaries, text patterns implemented as regular expressions, and rules that examine surrounding text. It does not classify papers through LLM prompts, use a trained document-layout model, or apply optical character recognition (OCR). Screening identifies papers that may contain an evaluation. Extraction then assigns model names, tasks, and evidence of execution separately. A model name can be sufficient to flag a paper for screening without being sufficient to count as an evaluation.

Table 28 describes the six stages. The distinction between them matters: recovering a name does not establish its task, and recovering a score does not establish that the authors produced it.

Table 28: The six extraction stages and the information each produces. Decisions follow the stated rules rather than independent review of every source passage.

Stage	Procedure
Recover text	Read stored LaTeX or retrieved source files. Convert HTML to text while preserving row and cell separators where available. Extract PDF text while retaining layout spacing. Scanned pages and image-only tables are not processed with OCR.
Screen papers	Search titles, abstracts, and body text for LLM terms and model aliases. Nearby evaluation terms determine signal strength. The body search checks up to twenty windows around matches and stores up to five excerpts. Text after the first recognized bibliography boundary is removed.
Identify models and tasks	Match family-specific name patterns and preserve release, size, and variant information. Assign tasks first through benchmark names, then task terms, and then title or abstract context if nearby evidence is insufficient. Unrecognized names or tasks may remain unresolved.
Read tables and prose	Split LaTeX tables into rows and cells. In HTML and PDF text, inspect numerical lines near model, metric, and task terms. In prose, search for evaluation or author-use statements near model names.
Determine use and eligibility	Infer execution from nearby table text, rerun statements, or model-use evidence elsewhere in the paper. Exclude detected copied scores, ambiguous names that do not refer to models, embedding-only identifiers, and selected supporting-role uses. Rules applied when records are merged can resolve some previously unclear table ownership.
Combine duplicate records	Expand supported task codes, standardize model names, and remove repeated paper–model–task records, retaining table evidence before prose evidence. Remove a general family name when a compatible specific model name appears in the same paper and family.

Differences between source formats. LaTeX tables are split at row and cell separators rather than interpreted as complete TeX documents. Nested macros, merged cells, headers, and numerical model names can be read incorrectly. HTML preserves some cell boundaries, whereas PDF extraction can merge columns. Results stored only as images are not recovered. Removing text after a bibliography boundary can exclude later appendices. Joining source files can also produce an order different from that of the compiled paper.

Determining whether authors ran a model. Evidence of execution can come from nearby text or from a model-specific statement elsewhere in the paper. Phrases indicating copied results help distinguish imported scores, but may occur beside results that the authors did produce. Citation handling also differs by format. A cited LaTeX row without an explicit rerun statement is restricted in performance extraction but may remain in the primary model–task data. PDF and HTML rows lack equivalent citation flags. Some prose evaluation statements qualify without a separate check of the outcome. The labels therefore record decisions from contextual rules, rather than independently confirmed execution.

Citing evaluation software does not by itself show whether the authors ran it or copied its reported scores. Only the former meets the intended inclusion criteria. Checking both the evaluation procedure and the origin of scores follows the reproducibility concerns discussed by Biderman et al. (2024).

K.2 Checks of the extracted records

The combined extraction contains 147,876 evidence records: 98,914 from tables or numerical lines and 48,962 from prose. Of 82,404 candidate records for the primary analysis, exclusion rules remove 789: 278 ambiguous-name cases, 452 supporting-role cases, 41 cases involving ambiguous use of "Phi," 12 copied-result cases, and six embedding or reranking identifiers. Assigning tasks and removing duplicates then reduces the remaining 81,615 evidence records to 65,863 paper–model–task records.

For each final record, the retained representative evidence comes from one of three formats: 21,313 LaTeX-table records, 12,575 PDF/HTML-line records, and 31,975 prose records. Diagnostic flags identify 1,997 records with a row citation but no explicit rerun statement; 8,895 for which table ownership was resolved when records were merged; and 26,982 prose records without a recognized metric term in the saved clause. These groups overlap. They are not counts of errors: a citation can identify software, and a qualitative outcome need not name a numerical metric.

Checks of the registers account for which papers were processed. Exclusion rules remove recognized ineligible uses, and duplicate removal prevents repeated evidence from increasing the count of the same evaluation. Appendix J examines how the results change under additional evidence requirements. Together, these checks make the calculations traceable to the extracted records.

K.3 Task assignment, source versions, and validation

Task assignment follows the sequence described in Appendix D: benchmark names, nearby task terms, then title or abstract context. Assignments are less specific when a benchmark mentioned for a whole table is applied to every row, or when the paper's topic supplies a model's task. Reapplying the same rules to nearby text tests dependence on context; it does not measure agreement with an expert.

Recovered text and available PDFs can come from different revisions. Excerpts, version information, and available file hashes help trace the records but do not establish that every source uses the same revision. First submission determines whether a paper enters the study, not when each extracted experiment was conducted.

The proposed numerical standards describe evaluations identified under these rules. Changes to model identification, inclusion decisions, or task assignment require recalculating the affected standards. The available checks support reproducible calculations from the extracted data; independent source review would address the separate question of classification accuracy.

L How the proposed standards are calculated and used

The proposed standards help researchers compare their model selections with observed evaluation practice. They combine numerical summaries of the papers studied with the selection principles in Section 3. They are starting points for explaining a design choice, not estimates of an optimal sample size or minimum requirements that make every evaluation valid. They also do not represent community consensus.

L.1 Selecting model, family, and task counts

For each contribution group, we calculate the 75th percentile separately for the number of models, model families, and task families. This places the proposed count above the median practice while reducing the influence of unusually large panels. We select an observed integer rather than interpolate between two counts. Given sorted observations $x_{(1)}, \ldots, x_{(n)}$, the rule is

$$R_q = x_{(1+\lceil (n-1)q \rceil)}, \qquad q = .75.$$

Here, n is the number of papers and $\lceil \cdot \rceil$ rounds upward to the next integer. We denote the resulting count by R_{75}. Because many papers can share the same count, more than 25% may reach the threshold. A paper can belong to more than one contribution group.

We draw 2,000 bootstrap samples of papers to calculate a 95% interval for each threshold. The reported model and family intervals extend by no more than one count from the estimated threshold. These intervals show how the thresholds vary when the extracted papers are resampled. They do not measure sampling error in a complete census or uncertainty from classifying source papers.

We keep the three counts separate because they serve different purposes. More releases can test variation within a model series; more families extend the range of series tested; and more tasks extend the settings in which a claim is examined. Combining them into one score would require deciding how much of one type of coverage compensates for less of another. Researchers should choose the relevant contribution group, then determine which kinds of coverage their claim requires.

L.2 How often papers meet several standards together

The separate 75th-percentile counts do not show how often a paper meets all of them. Table 29 calculates this directly. The first result column counts papers reaching their contribution group's model, family, and task thresholds together. The next also requires both open-weight and hosted models. The final column adds a model released within 170 days before first submission. Each column keeps the preceding conditions, and every percentage uses all papers in that row as its denominator. The recent model can appear on any task in the paper.

Table 29: Papers meeting several proposed standards together. *Breadth* requires all three listed model/family/task counts. Each subsequent column adds another condition: both access types, then a model released within 170 days before first submission. Percentages use the total paper count in the row. These are documented combinations of conditions, not compliance rates for a universal checklist.

Contribution	Papers	Models/families/tasks	Breadth	+ Open and hosted	+ Age $\leq$ 170 days
Overall reference	11,428	5/3/2	1,662 (14.5%)	1,104 (9.7%)	585 (5.1%)
New model or method	8,231	5/3/2	1,270 (15.4%)	872 (10.6%)	475 (5.8%)
Benchmark or evaluation	3,005	6/4/2	474 (15.8%)	396 (13.2%)	261 (8.7%)
Safety or alignment	3,198	5/3/2	574 (17.9%)	386 (12.1%)	198 (6.2%)
Domain application	1,744	5/4/3	167 (9.6%)	131 (7.5%)	68 (3.9%)
Training or adaptation	3,816	5/3/2	534 (14.0%)	330 (8.6%)	159 (4.2%)
Systems or efficiency	1,896	4/3/3	184 (9.7%)	81 (4.3%)	38 (2.0%)
Agents or interaction	1,789	5/3/2	298 (16.7%)	236 (13.2%)	150 (8.4%)

The overall thresholds of five models, three families, and two tasks are reached separately by 25.1%, 36.4%, and 48.8% of papers. Only 14.5% meet all three. Model and family counts are strongly related: their Spearman rank correlation is .879. Task counts are less closely related to model counts (.313) and family counts (.256). We therefore count the observed combinations rather than multiply the separate percentages. Across contribution groups, 9.6–17.9% meet all three count thresholds.

A separate calculation considers models evaluated on the same task. Across task families, 21.1–32.8% of paper–task combinations meet both the task's model-count and family-count thresholds. This calculation adds no access or recency requirement. It addresses coverage of a specific task rather than the three counts for an entire paper.

What missing release dates can change. At least one model can be dated to a release on or before first submission in 9,813 of 11,428 papers. All extracted models can be dated in 6,415 papers. The complete combination of counts, access types, and recency is documented in 585 papers: 5.1% of all papers, or 6.0% of papers with at least one dated model.

To examine missing information, we hold the extracted model, family, and task counts fixed and allow unmatched models to satisfy unresolved access or recency conditions. This gives an upper bound of 1,332 papers (11.7%). It bounds the effect of missing registry information, not errors in extracting evaluations. Appendix G explains how releases after first submission are handled.

These results show how much rarer it is to meet several conditions together. They are not full-compliance rates for Table 1, because the data do not identify every paper's central claim or all reporting requirements. Both access types are relevant when the claim spans them; recency is relevant when the claim concerns current capability. The research question determines which conditions should be combined.

L.3 Counts and release ages for specific tasks

For each paper p and task t, let a_{pt} be the age of the newest dated model evaluated on that task at first submission. We exclude unknown dates and releases after submission. The task's age threshold, $A_{25}(t)$, is the 25th percentile across these papers, using the observed-value rule above with $q = .25$. Each paper contributes once per task.

Table 30 reports task-specific thresholds, paper counts, and bootstrap intervals. Model and family counts use all papers assigned to the task; ages require a dated model on it. Ten task families have

thresholds of four models and three families. Systems has three models and two families; mathematics has four models and two families. Recommending three families therefore exceeds the observed 75th percentile for those two tasks. Family counts also do not establish independent model development (Appendix E).

Table 30: Proposed counts and release ages by task. *Papers* is the number used for model and family counts. *Dated* is the number with at least one model released on or before first submission on that task. Model and family thresholds are separate 75th percentiles. Age is the 25th percentile, $A_{25}(t)$, with a 95% paper-bootstrap interval. The row does not imply that a fixed percentage meets all thresholds together.

Task family	Papers	Identifiers	Families	Dated	Age, days [CI]
Agents, tools, and robotics	1,144	4	3	1,005	143 [127, 155]
Code and software engineering	1,372	4	3	1,223	148 [139, 156]
Knowledge and general reasoning	1,154	4	3	1,035	218 [193, 249]
Language and communication	2,651	4	3	2,299	195 [185, 207]
Mathematics and formal reasoning	1,726	4	2	1,513	188 [175, 209]
Medicine and health	1,583	4	3	1,395	154 [141, 165]
Models and systems	3,168	3	2	2,612	245 [224, 260]
Multimodal perception and generation	1,779	4	3	1,413	195 [185, 210]
Professional and social applications	1,707	4	3	1,453	153 [145, 164]
Retrieval, memory, and long context	1,426	4	3	1,271	188 [175, 203]
Safety, reliability, and society	2,236	4	3	1,944	168 [160, 176]
Science and discovery	479	4	3	389	113 [87, 143]

Age thresholds range from 113 to 245 days. They are lower for science, agents, and code than for general reasoning and systems. Overlapping intervals favor interpreting broad differences rather than a precise task ranking. Dated evidence ranges from 389 science papers to 2,612 systems papers. The overall 170-day threshold applies to paper-wide comparisons.

Using only models evaluated on a task gives age thresholds 12–43 days higher than using the newest model anywhere in the paper. The direction remains the same when both calculations use exactly the same papers, so the difference is not explained by which papers have known dates. In this same-paper comparison, the code threshold rises from 136 to 148 days and the systems threshold from 202 to 245 days. The numerical supplement includes these paired calculations. The overall 170-day threshold describes the newest model in each paper; it is not the percentile obtained by pooling all paper–task combinations.

L.4 Using and updating the standards

A table of models and tasks can show which models support each principal claim. Authors can compare its counts with the relevant standards and explain their choices: continuity with earlier work, differences in development, scale, release, or access type, or inclusion of a current competitor.

Updates should retain the counting units and percentile rule, document changes in dates, categories, source availability, and release records, and recalculate the separate thresholds and the proportions meeting them together. Future studies could test whether adding relevant families, releases, or access types changes a conclusion by more than a predefined amount. This would assess adequacy for specific claims rather than infer it from common practice.